Mathematics of Stochastic Processes & Randomness

Volume III

Random Fields, Point Processes & Spatial Randomness

Silvano Salvador

2026

Contents

Preface vii

I Functional-Analytic Foundations 1

1 Kernels, Hilbert Spaces, and the Kolmogorov Extension Theorem 2

1.1 Positive-Definite Kernels and Reproducing Kernel Hilbert Spaces . . . 2

1.1.1 Mercer's Theorem and Spectral Expansions . . . 3

1.2 The Kolmogorov Extension Theorem for Random Fields . . . 4

1.3 Stationarity and Isotropy . . . 5

1.4 Spectral Representations and Simulation of Stationary Fields . . . 6

1.5 Kernel Ridge Regression and the Representer Theorem . . . 8

2 Gaussian Processes and Fields: Foundations 10

2.1 Gaussian Vectors and the Multivariate Normal Distribution . . . 10

2.2 Gaussian Random Fields: Definition and Existence . . . 11

2.3 The Karhunen–Loève Expansion . . . 11

2.4 Sample Path Regularity: The Kolmogorov–Chentsov Theorem . . . 13

3 Concentration and Suprema of Gaussian Fields 15

3.1 The Fernique Theorem . . . 15

3.2 The Dudley Metric Entropy Bound . . . 15

3.3 Comparison Inequalities and Generic Chaining . . . 16

3.4 The Borell–TIS Inequality . . . 18

4 The Gaussian Free Field 19

4.1 Definition and Construction . . . 19

4.2 Conformal Invariance of the GFF . . . 20

4.3 Thick Points, Level Sets, and Liouville Quantum Gravity . . . 20

II Point Processes **22**

5 Poisson Point Processes **23**
5.1 Definition and Fundamental Properties 23
5.2 The Laplace Functional and Void Probabilities 24
5.3 The Campbell Theorem 25
5.4 Superposition, Thinning, and the Mapping Theorem 25
5.5 Palm Calculus 26

6 Cox Processes and Clustering **27**
6.1 Definition and Examples 27
6.2 Pair Correlation Functions and Second-Order Analysis 28
6.3 Ripley's K-Function and Summary Statistics 28

7 Gibbs Point Processes **29**
7.1 Interaction Potentials and the Gibbs Distribution 29
7.2 The DLR Equations and Infinite-Volume Gibbs Measures 30
7.3 MCMC Simulation of Gibbs Point Processes 30
7.4 Neyman–Scott Cluster Processes 31

III Spatial Gibbs Measures and Phase Transitions **33**

8 The Ising Model and Phase Transitions **34**
8.1 Definition of the Ising Model 34
8.2 The Peierls Argument: Phase Transition in $d \geq 2$ 35
8.3 The Lee–Yang Circle Theorem 35
8.4 FKG Inequality and Correlation Inequalities 36

IV Determinantal and Permanental Point Processes **37**

9 Determinantal Point Processes **38**
9.1 Definition and Characterisation 38
9.2 The Gaussian Unitary Ensemble and Eigenvalue Processes 39
9.3 Repulsiveness and the Pair Correlation Function 40
9.4 Sampling from DPPs 40

V Gaussian Processes on Metric Spaces and Manifolds **41**

10 Gaussian Processes Beyond $\mathbb{R}^d$ **42**
10.1 Kernels on Metric Spaces 42
10.2 Graph-Indexed Gaussian Processes and the Gaussian Markov Random Field 42
10.3 Gaussian Processes on Riemannian Manifolds 43

VI Spatial Statistics and Bayesian Nonparametrics **45**

11 Kriging and Spatial Interpolation **46**
11.1 The Kriging Predictor 46
11.2 Universal Kriging and the Intrinsic Model 46
11.3 Covariance Parameter Estimation by Maximum Likelihood 47

12 Gaussian Process Regression and Bayesian Nonparametrics **48**
12.1 Bayesian Inference with Gaussian Process Priors 48
12.2 Convergence of GP Regression: Posterior Contraction 49
12.3 Scalable Gaussian Process Methods 49
12.4 Variational Inference for Gaussian Processes 50

13 Spatial Statistics: Testing and Model Selection **52**
13.1 Testing for Complete Spatial Randomness 52
13.2 Variogram Estimation and Model Fitting 53
13.3 Spatial Prediction Under Model Uncertainty 53

VII Applications and Advanced Topics **54**

14 Applications in Physics, Biology, and Machine Learning **55**
14.1 Statistical Physics: Random Fields as Field Theories 55
14.2 Neuroscience and the Neural Field Model 55
14.3 Machine Learning: Deep Kernel Learning and Neural Networks as GPs 56

15 Stochastic Partial Differential Equations and Random Fields **57**
15.1 White Noise and the Sobolev Scale 57
15.2 Stochastic Reaction-Diffusion Equations 58

16 Worked Applications **59**
16.1 Kriging a Matérn Field: End-to-End Spatial Prediction 59

16.2 Fitting a Log-Gaussian Cox Process to Spatial Point Data 60
16.3 Ripley's K-Function: From PPP to Clustering and Regularity 62
16.4 The Gaussian Free Field: Discrete Approximation and Circle Average . 63
16.5 Determinantal Point Processes for Diverse Subset Selection 64

17 Permanental Point Processes and Bosonic Systems 67
17.1 Definition and Basic Properties 67
17.2 Cox Representation and Mixture Structure 68
17.3 Applications in Machine Learning: Repulsion versus Attraction 69

18 Random Matrices and Spatial Eigenvalue Distributions 70
18.1 The Wigner Semicircle Law 70
18.2 The Marchenko–Pastur Law for Sample Covariance Matrices 71
18.3 Beta Ensembles and Tridiagonal Models 72

19 Spatial Extremes and Max-Stable Processes 73
19.1 Extreme Value Theory for Spatial Fields 73
19.2 The Smith Model and the Brown–Resnick Process 74
19.3 Inference for Max-Stable Processes 75

20 Gaussian Processes on Graphs and Combinatorial Spaces 76
20.1 Heat Kernels on Graphs 76
20.2 Matérn-Type Kernels via Graph Differential Operators 77
20.3 Signal Processing on Graphs: Graph Fourier Transform 78

21 Gaussian Free Field: Further Topics 80
21.1 The Discrete GFF and the Random Walk Connection 80
21.2 Level Sets of the GFF and the Conformally Invariant Contours 81
21.3 The Gaussian Free Field in Higher Dimensions 82

22 Functional Data Analysis and Infinite-Dimensional Statistics 83
22.1 Functional Data and Covariance Operators 83
22.2 Estimation of the Covariance Operator 84
22.3 Functional Principal Components and FPCA 84
22.4 Functional Regression 85

23 Topological Data Analysis and Persistent Homology of Random Fields 86
23.1 Excursion Sets and Topological Complexity 86
23.2 Persistent Homology and the Persistence Diagram 87
23.3 Statistical Inference via the Persistence Diagram 87

24 Gaussian Processes in Reinforcement Learning and Bayesian Optimisation 89
24.1 Bayesian Optimisation 89
24.2 Thompson Sampling and the Exploration–Exploitation Trade-off 90
24.3 Multi-Fidelity and Multi-Task Gaussian Processes 91

25 Unified Perspectives: Kernels, Information, and Spatial Geometry 92
25.1 The Kernel as the Unifying Thread 92
25.2 Information-Theoretic Bounds and the Fisher Information Geometry of Random Fields 93
25.3 Spatial Geometry and the Intrinsic Random Field 94
25.4 From Random Fields to Random Measures: The Measure-Valued Perspective 95
25.5 Approximation Theory for Gaussian Processes: Nyström, Inducing Points, and Scalable Inference 95
25.6 Connections to Deep Learning: Neural Tangent Kernels and Feature Learning 97

Appendix A: Notation and Conventions 99

Appendix B: Summary of Main Theorems 100

Appendix C: Bibliographic Notes 102

Preface

This third volume completes the series *Mathematics of Stochastic Processes & Randomness* by developing the theory of randomness distributed over space. Where Volume I studied stochastic processes indexed by a discrete or continuous parameter $t \in \mathbb{N}$ or $t \in [0, \infty)$, and Volume II analysed the fine structure of paths through stochastic calculus, the present volume takes a genuinely multi-dimensional perspective: probability is spread over regions of $\mathbb{R}^d$, over graphs, over manifolds, over abstract measurable spaces carrying a geometric or topological structure.

The three principal objects of study are *random fields*, *point processes*, and *spatial models* arising from statistical physics and Bayesian nonparametrics. A *random field* is a collection $\{X_s : s \in T\}$ of random variables indexed by a parameter set T that is not merely a linearly ordered time axis but a multidimensional domain, a lattice, or an abstract metric space. A *point process* is a random measure whose realization consists of a (locally finite) collection of points scattered in some phase space; it models arrivals, events, particles, data points, and spatial patterns. The Gaussian free field, Gibbs measures, and Gaussian processes on manifolds round out the spatial picture.

The mathematical prerequisites are the contents of Volumes I and II together with a working knowledge of measure theory, functional analysis (Hilbert and Banach spaces, spectral theory of operators), and the rudiments of topology. The presentation follows the same template established in the preceding volumes: precise definitions, fully worked examples, complete proofs of central results, and algorithmic complements where applicable.

Part I lays the functional-analytic foundations: reproducing kernel Hilbert spaces, positive-definite kernels, and the Kolmogorov extension theorem in the spatial setting. Part II develops the theory of Gaussian random fields in depth, including sample-path regularity, the Fernique–Dudley–Borell framework, and the Gaussian free field. Part III treats point processes: Poisson processes in $\mathbb{R}^d$, Cox processes, Gibbs point processes, the Campbell theorem, and Palm calculus. Part IV studies spatial Gibbs measures and the Ising model, establishing phase transitions through the Peierls argument and the Lee–Yang theorem. Part V introduces determinantal and permanental point processes together with their connections to random matrix theory. Part VI covers Gaussian processes on metric spaces and manifolds, including Matérn covariances and their spectral representations. Part VII concludes with spatial statistics, kriging, and Bayesian nonparametric inference.

Part I

Functional-Analytic Foundations

Chapter 1

Kernels, Hilbert Spaces, and the Kolmogorov Extension Theorem

1.1 Positive-Definite Kernels and Reproducing Kernel Hilbert Spaces

The theory of random fields rests on a fundamental correspondence between positive-definite functions and Hilbert spaces of functions. This section develops the algebraic and analytic apparatus of reproducing kernel Hilbert spaces (RKHS), which will recur throughout the volume as the natural function-space setting for Gaussian processes.

Definition: Positive-Definite Kernel

Let T be a non-empty set. A function $k : T \times T \to \mathbb{R}$ is called a **positive-definite kernel** (or **covariance kernel**) if:

(i) **Symmetry**: $k(s,t) = k(t,s)$ for all $s, t \in T$;

(ii) **Positive semi-definiteness**: for every $n \geq 1$, every finite set of points $t_1, \ldots, t_n \in T$, and every $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \, c_j \, k(t_i, t_j) \geq 0.$$

Equivalently, every Gram matrix $K_{ij} = k(t_i, t_j)$ is positive semi-definite.

Positive-definiteness is the correct generalization of a covariance matrix to the infinite-dimensional setting. The central construction associated to each positive-definite kernel is a Hilbert space of functions on T, with the kernel serving simultaneously as the covariance and as the reproducing element.

Theorem: Moore–Aronszajn

Let $k : T \times T \to \mathbb{R}$ be a positive-definite kernel. There exists a unique Hilbert space $\mathcal{H}_k$ of real-valued functions on T such that:

(i) For each $t \in T$, the function $k(\cdot, t) \in \mathcal{H}_k$;

(ii) **Reproducing property**: for all $f \in \mathcal{H}_k$ and $t \in T$,

$$f(t) = \langle f, k(\cdot, t) \rangle_{\mathcal{H}_k}.$$

The space $\mathcal{H}_k$ is called the **reproducing kernel Hilbert space** (RKHS) associated to k.

Proof. **Existence.** Define the pre-Hilbert space $\mathcal{H}_0$ as the span of $\{k(\cdot, t) : t \in T\}$ with inner product determined by

$$\left\langle \sum_i \alpha_i k(\cdot, t_i), \sum_j \beta_j k(\cdot, s_j) \right\rangle = \sum_{i,j} \alpha_i \beta_j k(t_i, s_j).$$

Positive semi-definiteness of k ensures this is well-defined and non-negative. Completeness of the closure $\overline{\mathcal{H}_0}$ yields the RKHS $\mathcal{H}_k$. The reproducing property follows by definition of the inner product.

Uniqueness. Suppose $\mathcal{H}'$ is another RKHS with kernel k. For any $f \in \mathcal{H}'$, $f(t) = \langle f, k(\cdot, t) \rangle_{\mathcal{H}'}$, so $\mathcal{H}'$ and $\mathcal{H}_k$ agree on all evaluation functionals, and by density $\mathcal{H}' = \mathcal{H}_k$. □

Example: Squared-Exponential (RBF) Kernel

On $T = \mathbb{R}^d$, define

$$k_{\mathrm{SE}}(s, t) = \sigma^2 \exp\left(-\frac{\|s - t\|^2}{2\ell^2} \right)$$

with variance $\sigma^2 > 0$ and length-scale $\ell > 0$. This kernel is positive-definite (by Bochner's theorem, since its Fourier transform is a Gaussian density and hence non-negative), infinitely differentiable, and stationary. Its RKHS consists of infinitely smooth functions, and it is the canonical kernel for Gaussian processes in machine learning. The parameter ℓ controls the typical length of correlations: small ℓ yields rapidly fluctuating sample paths, large ℓ yields slowly varying ones.

1.1.1 Mercer's Theorem and Spectral Expansions

When T is a compact metric space equipped with a finite Borel measure μ, positive-definite kernels admit an eigenfunction expansion that diagonalises the associated integral operator.

Definition: Integral Operator of a Kernel

Let (T, μ) be a compact measure space and $k \in L^2(T \times T, \mu \otimes \mu)$ a positive-definite kernel. The **integral operator** $T_k : L^2(T, \mu) \to L^2(T, \mu)$ is defined by

$$(T_k f)(s) = \int_T k(s,t)\, f(t)\, d\mu(t).$$

Theorem: Mercer's Theorem

Let $T \subset \mathbb{R}^d$ be compact, μ a finite Borel measure, and $k : T \times T \to \mathbb{R}$ a continuous positive-definite kernel. Then T_k is a compact, self-adjoint, positive operator on $L^2(T, \mu)$ with non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ (counted with multiplicity) and orthonormal eigenfunctions $\{\phi_n\}_{n \geq 1}$. Moreover, the kernel admits the **Mercer expansion**

$$k(s,t) = \sum_{n=1}^{\infty} \lambda_n\, \phi_n(s)\, \phi_n(t),$$

where the series converges absolutely and uniformly on $T \times T$.

The Mercer expansion gives a canonical orthonormal basis for the RKHS: an element $f \in \mathcal{H}_k$ can be written as $f = \sum_n a_n \phi_n$ where $\sum_n a_n^2/\lambda_n < \infty$, and the RKHS norm is $\|f\|_{\mathcal{H}_k}^2 = \sum_n a_n^2/\lambda_n$.

1.2 The Kolmogorov Extension Theorem for Random Fields

A random field is defined by its finite-dimensional distributions. The Kolmogorov extension theorem gives a precise condition under which such a consistent family determines a genuine probability measure on a function space.

Definition: Random Field

A **random field** indexed by T is a collection of random variables $\{X_t : t \in T\}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in a measurable space $(E, \mathcal{E})$. The **finite-dimensional distributions** (fidi's) are the laws

$$\mu_{t_1,\ldots,t_n} = \mathcal{L}(X_{t_1}, \ldots, X_{t_n}), \qquad n \geq 1,\ t_1, \ldots, t_n \in T.$$

Theorem: Kolmogorov Extension

Let T be an arbitrary index set and $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Suppose we are given a family of probability measures $\{\mu_{t_1,\ldots,t_n}\}$ on $\mathbb{R}^n$, indexed by all finite subsets $\{t_1, \ldots, t_n\} \subset T$, satisfying:

(i) **Symmetry**: $\mu_{t_{\pi(1)},\ldots,t_{\pi(n)}} = \mu_{t_1,\ldots,t_n} \circ \pi^{-1}$ for every permutation π;

(ii) **Consistency**: $\mu_{t_1,\ldots,t_n}(A \times \mathbb{R}) = \mu_{t_1,\ldots,t_{n-1}}(A)$ for all $A \in \mathcal{B}(\mathbb{R}^{n-1})$.

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random field $\{X_t : t \in T\}$ with $\Omega = \mathbb{R}^T$ (the product space) such that the fidi's of $\{X_t\}$ coincide with the given family.

Remark: Regularity of Sample Paths

The Kolmogorov extension theorem provides existence on the product space $\mathbb{R}^T$, but sample paths $t \mapsto X_t(\omega)$ are in general merely measurable functions, not continuous. Establishing continuity or Hölder regularity requires additional hypotheses on the covariance structure, as developed in Chapter 3 (Kolmogorov–Chentsov theorem for fields).

1.3 Stationarity and Isotropy

The most tractable class of random fields enjoys spatial homogeneity: the statistical properties are invariant under translations and rotations of the domain.

Definition: Stationarity and Isotropy

A random field $\{X_s : s \in \mathbb{R}^d\}$ with $\mathbb{E}[X_s^2] < \infty$ for all s is called:

- **(Weakly) stationary** if $\mathbb{E}[X_s] = m$ is constant and $\operatorname{Cov}(X_s, X_t) = C(s - t)$ depends only on the displacement $s - t$;
- **Isotropic** if additionally $C(s - t) = \rho(\|s - t\|)$ depends only on the Euclidean distance $\|s - t\|$.

The function $C : \mathbb{R}^d \to \mathbb{R}$ is called the **covariance function** and $\rho : [0, \infty) \to \mathbb{R}$ the **radial covariance** or **variogram**.

Theorem: Bochner's Theorem

A continuous function $C : \mathbb{R}^d \to \mathbb{R}$ with $C(0) > 0$ is the covariance function of a weakly stationary random field if and only if it is the Fourier transform of a finite non-negative Borel measure ν on $\mathbb{R}^d$:

$$C(h) = \int_{\mathbb{R}^d} e^{i\langle \xi, h \rangle} \, d\nu(\xi).$$

The measure ν is called the **spectral measure** of the field.

Proof. **Sufficiency.** If $C = \hat{\nu}$, then for any finite set of points $\{s_j\}$ and complex coefficients $\{c_j\}$,

$$\sum_{j,k} c_j \overline{c_k} C(s_j - s_k) = \int \left| \sum_j c_j e^{i\langle \xi, s_j \rangle} \right|^2 d\nu(\xi) \geq 0,$$

confirming positive-definiteness.

Necessity. A continuous positive-definite function on $\mathbb{R}^d$ is the Fourier transform of a finite positive measure by the Bochner–Schwartz theorem applied to tempered distributions. $\square$

Example: Matérn Covariance Family

The **Matérn class** of isotropic covariances on $\mathbb{R}^d$ is parameterised by $\nu > 0$ (smoothness) and $\ell > 0$ (length-scale):

$$\rho_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, r}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\, r}{\ell} \right),$$

where K_ν is the modified Bessel function of the second kind. Key special cases:

- $\nu = 1/2$: $\rho(r) = e^{-r/\ell}$ (Ornstein–Uhlenbeck, continuous but not differentiable);
- $\nu = 3/2$: $\rho(r) = (1 + \sqrt{3}r/\ell)e^{-\sqrt{3}r/\ell}$ (once differentiable);
- $\nu \to \infty$: $\rho(r) \to e^{-r^2/(2\ell^2)}$ (squared-exponential, infinitely differentiable).

A Gaussian field with Matérn-ν covariance has sample paths that are $(k-1)$-times differentiable when $\nu > k - 1$, with $\lfloor \nu \rfloor$ being the order of mean-square differentiability.

1.4 Spectral Representations and Simulation of Stationary Fields

Bochner's theorem guarantees that every stationary covariance admits a spectral measure, but for simulation and computation one needs explicit formulas. When the spectral measure has a density $S(\xi)$ (the **spectral density**), the field admits a stochastic integral representation that is particularly amenable to Fast Fourier Transform (FFT)-based simulation.

Definition: Spectral Density

If the spectral measure ν is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$, the density $S : \mathbb{R}^d \to [0, \infty)$ satisfying $d\nu(\xi) = S(\xi)d\xi$ is called the **spectral density** or **power spectral density** of the field. By Bochner's theorem,

$$C(h) = \int_{\mathbb{R}^d} e^{i\langle \xi, h \rangle} S(\xi)\, d\xi = \hat{S}(h),$$

so the covariance function and spectral density are a Fourier pair.

Example: Spectral Densities of Common Kernels

For isotropic kernels on $\mathbb{R}^d$, the spectral density depends only on $\|\xi\|$.
Squared-exponential. $C(h) = \sigma^2 e^{-\|h\|^2/(2\ell^2)}$ has spectral density

$$S(\xi) = \sigma^2 (2\pi\ell^2)^{d/2} e^{-2\pi^2\ell^2\|\xi\|^2}.$$

The spectrum is also Gaussian, reflecting the self-dual nature of the squared-exponential under Fourier transform.
Matérn-ν. The spectral density is

$$S(\xi) = \sigma^2 \cdot \frac{2^d \pi^{d/2} \Gamma(\nu + d/2)(2\nu/\ell^2)^\nu}{\Gamma(\nu)} \cdot \left(\frac{2\nu}{\ell^2} + 4\pi^2\|\xi\|^2\right)^{-(\nu+d/2)},$$

a heavy-tailed power law with exponent $\nu + d/2$. The slower decay at high frequencies (for small ν) corresponds to the rougher sample paths.

Theorem: Spectral Representation of a Stationary Gaussian Field

Let $X \sim \mathcal{GP}(0, C)$ be a mean-zero stationary Gaussian field on $\mathbb{R}^d$ with spectral density S. Then X has the stochastic integral representation

$$X(s) = \int_{\mathbb{R}^d} e^{i\langle \xi, s\rangle}\, \tilde{\mathcal{W}}(d\xi),$$

where $\tilde{\mathcal{W}}$ is a **complex Gaussian random measure** with $\mathbb{E}[\tilde{\mathcal{W}}(A)\overline{\tilde{\mathcal{W}}(B)}] = \int_{A\cap B} S(\xi)\, d\xi$ and the conjugate symmetry $\tilde{\mathcal{W}}(-A) = \overline{\tilde{\mathcal{W}}(A)}$ (ensuring $X(s)$ is real).

FFT-based simulation on a grid

For simulation on a d-dimensional grid $\{0, h, 2h, \ldots, (n-1)h\}^d$ with $N = n^d$ total points, the spectral representation yields a simulation algorithm of cost $O(N \log N)$:

1. Compute the spectral density values $S(\xi_k)$ at the Fourier frequencies $\xi_k = k/(nh)$ for $k \in \{0, 1, \ldots, n-1\}^d$;
2. Draw independent complex Gaussian values $Z_k \sim \mathcal{N}_{\mathbb{C}}(0, 1)$ for each frequency;
3. Set $\tilde{X}_k = \sqrt{S(\xi_k)} \cdot Z_k$ (conjugate-symmetric for real output);
4. Compute $X = \mathrm{IFFT}(\tilde{X})$, the inverse discrete Fourier transform.

The resulting values $(X(s_j))$ have exactly the desired covariance structure on the grid, provided the grid is large enough that the circular boundary condition imposed by the FFT does not introduce artefacts. Embedding the target grid in a larger grid (typically $2\times$ in each dimension) and discarding the boundary eliminates the wrap-around effect.

Remark: Circulant Embedding for Non-Negative Definite Embeddings

For the FFT simulation to work, the spectral density values $S(\xi_k)$ must all be non-negative. This is guaranteed for any valid covariance function. However, the **circulant embedding** of a covariance matrix (wrapping the covariance on the torus) may fail to be positive semi-definite for small grids; the fix is to embed into a larger grid until positivity is achieved. For Matérn and squared-exponential kernels, an embedding factor of 2 is typically sufficient in practice.

1.5 Kernel Ridge Regression and the Representer Theorem

The reproducing kernel Hilbert space machinery developed above is not merely a tool for describing covariance structures: it underlies a general method for nonparametric estimation in which a regularised optimisation problem over an infinite-dimensional function space reduces, by a single structural theorem, to a finite-dimensional linear algebra problem of size n (the sample size). This is the content of the representer theorem, and it explains why the RKHS formalism recurs throughout this volume in contexts that are ostensibly purely probabilistic.

Definition: Kernel Ridge Regression

Let k be a positive-definite kernel on T with RKHS $\mathcal{H}_k$, and let $(x_1, y_1), \dots, (x_n, y_n) \in T \times \mathbb{R}$ be observed data. For a regularisation parameter $\lambda > 0$, the **kernel ridge regression** estimator is the minimiser of the penalised empirical risk

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{H}_k} J(f), \qquad J(f) = \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \, \|f\|_{\mathcal{H}_k}^2 .$$

A priori, the minimisation in the definition above ranges over an infinite-dimensional space, and there is no reason to expect a closed-form solution. The representer theorem shows that the minimiser always lies in the finite-dimensional subspace spanned by the kernel sections at the observed points.

Theorem: Representer Theorem

The minimiser $\hat{f}$ of J in $\mathcal{H}_k$ exists, is unique, and admits the representation

$$\hat{f}(\cdot) = \sum_{i=1}^{n} c_i \, k(\cdot, x_i)$$

for some coefficient vector $\mathbf{c} = (c_1, \dots, c_n)^\top \in \mathbb{R}^n$.

Proof. Let $V = \operatorname{span}\{k(\cdot, x_1), \dots, k(\cdot, x_n)\} \subset \mathcal{H}_k$, a finite-dimensional (hence closed) subspace, and write any $f \in \mathcal{H}_k$ as the orthogonal decomposition $f = f_{\parallel} + f_{\perp}$ with

$f_\parallel \in V$ and $f_\perp \in V^\perp$. By the reproducing property, for each i,

$$f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k} = \langle f_\parallel, k(\cdot, x_i)\rangle_{\mathcal{H}_k} + \langle f_\perp, k(\cdot, x_i)\rangle_{\mathcal{H}_k} = f_\parallel(x_i),$$

since $k(\cdot, x_i) \in V$ and $f_\perp \perp V$. Hence the data-fit term $\sum_i (y_i - f(x_i))^2$ depends on f only through $f_\parallel$, and is unaffected by $f_\perp$. Meanwhile, by orthogonality, $\|f\|_{\mathcal{H}_k}^2 = \|f_\parallel\|_{\mathcal{H}_k}^2 + \|f_\perp\|_{\mathcal{H}_k}^2 \geq \|f_\parallel\|_{\mathcal{H}_k}^2$, with equality iff $f_\perp = 0$. Therefore $J(f) \geq J(f_\parallel)$ for every f, with equality iff $f_\perp = 0$, so the minimiser satisfies $\hat{f} = \hat{f}_\parallel \in V$, which is precisely the stated representation. Strict convexity of J on V (as $\lambda > 0$ and K is positive semi-definite) gives uniqueness. □

Corollary: Closed-Form Solution

Substituting $\hat{f} = \sum_i c_i k(\cdot, x_i)$ into J, using $\hat{f}(x_j) = \sum_i c_i K_{ij}$ and $\left\|\hat{f}\right\|_{\mathcal{H}_k}^2 = \mathbf{c}^\top K \mathbf{c}$ (where $K_{ij} = k(x_i, x_j)$), reduces J to the finite-dimensional quadratic

$$J(\mathbf{c}) = \|\mathbf{y} - K\mathbf{c}\|^2 + \lambda\, \mathbf{c}^\top K \mathbf{c},$$

whose minimiser, obtained by setting $\nabla_{\mathbf{c}} J = 0$, is

$$\hat{\mathbf{c}} = (K + \lambda I)^{-1}\mathbf{y}.$$

Remark: Equivalence with the Gaussian Process Posterior Mean

Comparing the closed-form solution above with the posterior mean of Gaussian process regression (Chapter 12), $m_{\text{post}}(x^*) = k(x^*, \mathbf{x})(K + \sigma_n^2 I)^{-1}\mathbf{y}$, shows that the kernel ridge regression predictor $\hat{f}(x^*) = \sum_i \hat{c}_i k(x^*, x_i) = k(x^*, \mathbf{x})(K + \lambda I)^{-1}\mathbf{y}$ coincides *exactly* with the GP posterior mean once the regularisation parameter is identified with the observation noise variance, $\lambda = \sigma_n^2$. This is the classical correspondence between regularised function estimation (a frequentist, variational viewpoint) and Bayesian nonparametric regression with a Gaussian process prior: the RKHS norm penalty $\lambda \|f\|_{\mathcal{H}_k}^2$ plays exactly the role of the negative log-density of the GP prior, restricted to the relevant finite-dimensional projection. The posterior *variance*, however, has no analogue on the regularisation side: it is a genuinely Bayesian quantity, reflecting uncertainty rather than merely a point estimate.

Chapter 2

Gaussian Processes and Fields: Foundations

2.1 Gaussian Vectors and the Multivariate Normal Distribution

Definition: Gaussian Random Vector

A random vector $\mathbf{X} = (X_1, \dots, X_n)^\top$ taking values in $\mathbb{R}^n$ is **Gaussian** (or **multivariate normal**) if every linear combination $\sum_{i=1}^n a_i X_i$ is a (univariate) Gaussian random variable. Its distribution is characterised by:

- Mean vector $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] \in \mathbb{R}^n$;
- Covariance matrix $\Sigma_{ij} = \operatorname{Cov}(X_i, X_j)$, which is symmetric positive semi-definite.

We write $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

When Σ is strictly positive definite, the density of $\mathbf{X}$ with respect to Lebesgue measure on $\mathbb{R}^n$ is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} (\det \Sigma)^{1/2}} \exp\!\left(-\tfrac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

The moment-generating function is $M(\mathbf{u}) = \exp(\boldsymbol{\mu}^\top \mathbf{u} + \tfrac{1}{2} \mathbf{u}^\top \Sigma \mathbf{u})$, and all cumulants of order ≥ 3 vanish.

Proposition: Conditional Distribution of a Gaussian Vector

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with block structure

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

If Σ_{22} is invertible, the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is Gaussian:

$$\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2 \;\sim\; \mathcal{N}\!\left(\boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2),\; \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right).$$

This conditional Gaussian formula is the key to kriging interpolation (Chapter 12) and Gaussian process regression (Chapter 13).

2.2 Gaussian Random Fields: Definition and Existence

Definition: Gaussian Random Field

A random field $\{X_t : t \in T\}$ is a **Gaussian random field** (GRF) if every finite-dimensional marginal $(X_{t_1}, \dots, X_{t_n})$ is multivariate Gaussian. A GRF is completely characterised by its:

- **Mean function**: $m(t) = \mathbb{E}[X_t]$;
- **Covariance kernel**: $k(s,t) = \operatorname{Cov}(X_s, X_t)$.

We write $X \sim \mathcal{GP}(m, k)$.

Theorem: Existence of Gaussian Random Fields

For any function $m : T \to \mathbb{R}$ and any positive-definite kernel $k : T \times T \to \mathbb{R}$, there exists a Gaussian random field $X \sim \mathcal{GP}(m, k)$.

Proof. The finite-dimensional marginals $\mathcal{N}(\boldsymbol{m}_{\mathbf{t}}, K_{\mathbf{t}})$, where $\boldsymbol{m}_{\mathbf{t}} = (m(t_i))$ and $(K_{\mathbf{t}})_{ij} = k(t_i, t_j)$, form a consistent family (Gaussian marginals close under conditioning and marginalization). The Kolmogorov Extension Theorem provides a probability measure on $\mathbb{R}^T$ with these marginals. □

2.3 The Karhunen–Loève Expansion

Mercer's theorem (Chapter 1) expands a covariance kernel in terms of its eigenfunctions; the Karhunen–Loève theorem is its probabilistic counterpart, expanding the *random field itself* as a series in those same eigenfunctions with independent Gaussian coefficients. This expansion is the basic tool for simulating Gaussian fields on compact domains, for truncating infinite-dimensional fields to finite-dimensional approximations with explicit, optimal error control, and for understanding the sample-path structure of the field through the decay of the kernel's spectrum.

Theorem: Karhunen–Loève Expansion

Let $X \sim \mathcal{GP}(0, k)$ on a compact metric space T equipped with a finite Borel measure μ, where k is continuous and $\int_T k(t,t)\, d\mu(t) < \infty$. Let $(\lambda_n, \phi_n)_{n \geq 1}$ be the Mercer eigenpairs of k, as in Mercer's theorem (Chapter 1). Then X admits the representation

$$X(t) = \sum_{n=1}^{\infty} \sqrt{\lambda_n}\, \xi_n\, \phi_n(t), \qquad t \in T,$$

where $\xi_n = \lambda_n^{-1/2} \int_T X(s)\phi_n(s)\,d\mu(s)$ are i.i.d. $\mathcal{N}(0,1)$ random variables, and the series converges in $L^2(\mathbb{P}\otimes\mu)$, and uniformly in t almost surely whenever $\sum_n \lambda_n < \infty$.

Proof. Define ξ_n as in the statement; since X is Gaussian and the ϕ_n are orthonormal in $L^2(T,\mu)$, the vector $(\xi_1,\dots,\xi_N)$ is jointly Gaussian for every N, with

$$\operatorname{Cov}(\xi_m,\xi_n) = (\lambda_m\lambda_n)^{-1/2}\int_T\int_T k(s,t)\phi_m(s)\phi_n(t)\,d\mu(s)\,d\mu(t) = (\lambda_m\lambda_n)^{-1/2}\lambda_n\delta_{mn} = \delta_{mn},$$

using the eigenfunction equation $\int_T k(s,t)\phi_n(t)\,d\mu(t) = \lambda_n\phi_n(s)$ and orthonormality of $\{\phi_n\}$. Hence the ξ_n are i.i.d. standard Gaussian. For the convergence, set $S_N(t) = \sum_{n=1}^{N}\sqrt{\lambda_n}\xi_n\phi_n(t)$; by the Mercer expansion $k(t,t) = \sum_n \lambda_n\phi_n(t)^2$, so

$$\mathbb{E}\big[(X(t)-S_N(t))^2\big] = k(t,t) - \sum_{n=1}^{N}\lambda_n\phi_n(t)^2 \to 0$$

pointwise as $N\to\infty$, and integrating over $t\in T$ with respect to μ gives $L^2(\mathbb{P}\otimes\mu)$ convergence by dominated convergence, using $\int_T k(t,t)\,d\mu(t) < \infty$. Uniform almost-sure convergence under $\sum_n\lambda_n<\infty$ follows from Mercer's theorem's uniform convergence of $\sum_n\lambda_n\phi_n(s)\phi_n(t)$ together with a martingale argument applied to the tail sums $\sum_{n>N}\sqrt{\lambda_n}\xi_n\phi_n(t)$. □

Corollary: Optimality of the Truncated Expansion

Among all rank-N approximations $\hat{X}_N(t) = \sum_{n=1}^{N} a_n(t)Z_n$ with Z_n orthonormal random variables and $a_n \in L^2(T,\mu)$, the Karhunen–Loève truncation S_N minimises the total mean-squared error $\int_T \mathbb{E}[(X(t)-\hat{X}_N(t))^2]\,d\mu(t)$, with minimal value $\sum_{n>N}\lambda_n$.

Example: Karhunen–Loève Expansion of Brownian Motion

For Brownian motion on $T=[0,1]$ with $k(s,t)=\min(s,t)$, the Mercer eigenpairs are explicit:

$$\lambda_n = \frac{1}{\big(n-\tfrac12\big)^2\pi^2}, \qquad \phi_n(t) = \sqrt{2}\,\sin\!\big((n-\tfrac12)\pi t\big), \qquad n=1,2,\dots,$$

so that

$$W_t = \sqrt{2}\sum_{n=1}^{\infty}\xi_n\,\frac{\sin\big((n-\tfrac12)\pi t\big)}{(n-\tfrac12)\pi}, \qquad \xi_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

This is the classical **Lévy–Ciesielski construction** of Brownian motion (in its Karhunen–Loève rather than Haar-wavelet form). Since $\lambda_n = O(n^{-2})$, the truncation error $\sum_{n>N}\lambda_n = O(1/N)$ decays only polynomially, reflecting the relatively rough ($\nu=1/2$, Hölder-$\tfrac12^-$) sample paths of Brownian motion; for the smoother squared-exponential kernel, by contrast, the Mercer eigenvalues de-

cay super-exponentially, and truncations with N as small as a few dozen terms typically suffice for practical simulation accuracy.

Remark: Relation to the Gaussian Free Field Series

The series representation of the Gaussian free field given in Chapter 4, $\Phi = \sum_n \xi_n e_n/\sqrt{\lambda_n}$ with (λ_n, e_n) the Dirichlet eigenpairs of $-\Delta$, is formally the Karhunen–Loève expansion of Φ with respect to the kernel G_D, except that the relevant "eigenvalues" of G_D are $1/\lambda_n$ (the Green's function is the inverse of the Laplacian) and the series converges only in the negative Sobolev space $H^{-\varepsilon}(D)$ rather than in $L^2(D,\mu)$, precisely because $\sum_n 1/\lambda_n = \infty$ in two dimensions: Weyl's law gives $\lambda_n \asymp n$ as $n \to \infty$, so $\sum_n 1/\lambda_n \asymp \sum_n 1/n$ diverges (logarithmically). This is the analytic origin of the GFF's status as a distribution rather than a function.

2.4 Sample Path Regularity: The Kolmogorov–Chentsov Theorem

The Kolmogorov Extension Theorem guarantees the existence of a random field on the product space $\mathbb{R}^T$, but gives no information about the regularity of sample paths $t \mapsto X_t(\omega)$. For $T \subset \mathbb{R}^d$, the following theorem gives Hölder continuity of paths from moment conditions.

Theorem: Kolmogorov–Chentsov for Random Fields

Let $\{X_t : t \in [0,1]^d\}$ be a random field. Suppose there exist constants $\alpha > 0$, $\beta > d$, $C > 0$ such that for all $s, t \in [0,1]^d$,

$$\mathbb{E}[|X_s - X_t|^\alpha] \leq C \, \|s - t\|^\beta .$$

Then X has a modification $\tilde{X}$ with **Hölder continuous** sample paths: there exists a random variable $H < \infty$ a.s. such that

$$\left|\tilde{X}_s - \tilde{X}_t\right| \leq H \, \|s - t\|^\gamma \qquad \forall\, s, t \in [0,1]^d,$$

for any $\gamma \in (0, (\beta - d)/\alpha)$.

Proof. The proof extends the one-dimensional Kolmogorov–Chentsov argument using dyadic cubes in $[0,1]^d$. Define the increments on the dyadic grid $\mathcal{D}_n = \{k \cdot 2^{-n} : k \in \{0, \ldots, 2^n\}^d\}$ at scale n. By Markov's inequality and the moment condition,

$$\mathbb{P}\left(\sup_{s,t \in \mathcal{D}_n, \|s-t\| \leq \sqrt{d}\cdot 2^{-n}} |X_s - X_t| > 2^{-n\gamma} \right) \leq C' \cdot 2^{nd} \cdot 2^{-n(\beta - d\alpha^{-1}\cdot\alpha)} = C' \cdot 2^{-n(\beta - d - \gamma\alpha)}.$$

Choosing $\gamma < (\beta - d)/\alpha$, the Borel–Cantelli lemma gives the result on dyadic grids, and continuity extends to $[0,1]^d$ by density. □

Corollary: Hölder Regularity of Gaussian Fields

If $X \sim \mathcal{GP}(m, k)$ with k satisfying $k(t,t) - k(s,t) - k(t,s) + k(s,s) \leq C \|s - t\|^{2\alpha}$ for some $\alpha > 0$, then X has a modification with Hölder continuous paths of exponent $\gamma < \alpha$ in all spatial directions.

Chapter 3

Concentration and Suprema of Gaussian Fields

3.1 The Fernique Theorem

A fundamental result in the theory of Gaussian measures is that the norm of a Gaussian random variable in a Banach space has Gaussian tails. For a Gaussian field, this translates into control over the supremum.

Theorem: Fernique

Let $(X_t)_{t\in T}$ be a separable Gaussian random field with $\mathbb{E}[\sup_t X_t] < \infty$. Then there exists $\alpha > 0$ such that

$$\mathbb{E}\left[\exp\left(\alpha \left(\sup_{t\in T} X_t\right)^2\right)\right] < \infty.$$

In particular, $\sup_{t\in T} X_t$ has Gaussian tails: $\mathbb{P}(\sup_t X_t > u) \le Ce^{-\alpha u^2}$ for all u large enough.

Remark: Integrability of All Moments

Fernique's theorem implies that all moments $\mathbb{E}[(\sup_t X_t)^p] < \infty$ for $p \ge 1$. This is far stronger than what holds for general processes: for a Gaussian field, the supremum is essentially sub-Gaussian.

3.2 The Dudley Metric Entropy Bound

Let $d_X(s,t) = (\mathbb{E}[(X_s - X_t)^2])^{1/2}$ be the **canonical metric** of the Gaussian field (X_t). The Dudley bound controls $\mathbb{E}[\sup_t X_t]$ through the metric entropy of (T, d_X).

Definition: Metric Entropy

For a metric space (T, d) and $\varepsilon > 0$, the **covering number** $N(T, d, \varepsilon)$ is the smallest number of balls of radius ε needed to cover T. The function $\varepsilon \mapsto \log N(T, d, \varepsilon)$ is called the **metric entropy** of (T, d).

Theorem: Dudley's Entropy Bound

Let $(X_t)_{t \in T}$ be a mean-zero separable Gaussian random field. Then

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \lesssim \int_0^{\operatorname{diam}(T, d_X)} \sqrt{\log N(T, d_X, \varepsilon)}\, d\varepsilon.$$

Example: Brownian Sheet

The **Brownian sheet** $\{W_t : t \in [0,1]^d\}$ is a mean-zero Gaussian field with covariance

$$k(s, t) = \prod_{i=1}^{d} \min(s_i, t_i).$$

Its canonical metric satisfies $d_W(s,t)^2 = \prod_i s_i + \prod_i t_i - 2\prod_i \min(s_i, t_i)$, and the Dudley bound gives $\mathbb{E}[\sup_{t \in [0,1]^d} W_t] = O((\log n)^{d/2})$ as the grid refines. Paths of the Brownian sheet are continuous but have non-differentiable structure in each coordinate direction individually.

3.3 Comparison Inequalities and Generic Chaining

The Dudley bound of the preceding section controls $\mathbb{E}[\sup_t X_t]$ from above using a single scalar summary of the metric entropy, but it is sometimes loose by logarithmic factors. Two complementary refinements sharpen the picture: comparison theorems, which bound the supremum of one Gaussian field by that of another with larger increments, and the generic chaining method, which replaces entropy numbers by a genuinely two-sided, sharp functional of the metric space.

Proposition: Slepian's Inequality

Let $X = (X_t)_{t \in T}$ and $Y = (Y_t)_{t \in T}$ be centred Gaussian fields on a finite index set T with $\mathbb{E}[X_t^2] = \mathbb{E}[Y_t^2]$ for all t, and suppose

$$\mathbb{E}\big[(X_s - X_t)^2\big] \le \mathbb{E}\big[(Y_s - Y_t)^2\big] \qquad \text{for all } s, t \in T.$$

Then $\mathbb{E}[\sup_t X_t] \le \mathbb{E}[\sup_t Y_t]$.

Proof. By rescaling we may assume X and Y are defined on the same probability space and interpolate via the **Gaussian smart path**: for $\theta \in [0,1]$, let $Z_\theta = \sqrt{\theta}\, Y + \sqrt{1-\theta}\, X'$, where X' is an independent copy of X with the same covariance, so that Z_θ has co-

variance $\theta\,\mathrm{Cov}(Y)+(1-\theta)\,\mathrm{Cov}(X)$ and interpolates continuously between $\mathrm{Cov}(X)$ at $\theta=0$ and $\mathrm{Cov}(Y)$ at $\theta=1$. For a smooth function F (a suitable smoothed maximum), the Gaussian integration-by-parts (Stein's) identity gives

$$\frac{d}{d\theta}\mathbb{E}[F(Z_\theta)] = \frac{1}{2}\sum_{s,t}\Big(\mathrm{Cov}(Y_s,Y_t)-\mathrm{Cov}(X_s,X_t)\Big)\,\mathbb{E}\Big[\frac{\partial^2 F}{\partial z_s\partial z_t}(Z_\theta)\Big].$$

For F a smooth approximation of $\max_t z_t$, the off-diagonal second partials of F are non-positive (the maximum is a "submodular"-type function in this sense), and the hypothesis $\mathbb{E}[(X_s-X_t)^2]\le\mathbb{E}[(Y_s-Y_t)^2]$ together with equal variances translates into $\mathrm{Cov}(Y_s,Y_t)\ge\mathrm{Cov}(X_s,X_t)$ for $s\neq t$, so each term in the sum is non-negative. Hence $\theta\mapsto\mathbb{E}[F(Z_\theta)]$ is non-decreasing, giving $\mathbb{E}[F(X)]\le\mathbb{E}[F(Y)]$; letting the smoothing parameter of F tend to the true maximum completes the proof. □

Remark: Sudakov's Lower Bound

A converse-type bound, dual to Dudley's upper bound, follows from Slepian's inequality applied to a well-separated finite subset: if there exist N points $t_1,\dots,t_N\in T$ with pairwise canonical distances $d_X(t_i,t_j)\ge\delta$ for all $i\neq j$, then comparing X restricted to $\{t_1,\dots,t_N\}$ with N i.i.d. $\mathcal{N}(0,\delta^2/2)$ variables (which have smaller pairwise increments) gives, via Slepian,

$$\mathbb{E}\Big[\sup_{1\le i\le N} X_{t_i}\Big]\ge\frac{\delta}{2}\sqrt{\log N}.$$

This is **Sudakov's minoration**: $\mathbb{E}[\sup_t X_t]\gtrsim\sup_{\delta>0}\delta\sqrt{\log N(T,d_X,\delta)}$, the same metric-entropy quantity appearing in Dudley's bound but now evaluated at a single scale rather than integrated, and providing a matching lower bound up to the choice of scale.

Theorem: Talagrand's Generic Chaining (Majorizing Measures)

For a centred Gaussian field $(X_t)_{t\in T}$ with canonical metric d_X, define the **generic chaining functional**

$$\gamma_2(T,d_X)=\inf_{(\mathcal{A}_n)}\sup_{t\in T}\sum_{n\ge0}2^{n/2}d_X(t,A_n(t)),$$

where the infimum is over all admissible sequences of partitions $\mathcal{A}_n$ of T with $|\mathcal{A}_0|=1$ and $|\mathcal{A}_n|\le 2^{2^n}$, and $A_n(t)$ denotes the unique element of $\mathcal{A}_n$ containing t. Then

$$\frac{1}{L}\,\gamma_2(T,d_X)\le\mathbb{E}\Big[\sup_{t\in T}X_t\Big]\le L\,\gamma_2(T,d_X)$$

for a universal constant $L>0$, independent of T, d_X, and the field.

Remark: Why Chaining Sharpens Dudley

Dudley's bound is recovered from γ_2 by using the same partition cardinalities to bound entropy numbers, and is generally a valid upper bound on γ_2 itself; the two coincide up to universal constants whenever the metric space (T, d_X) is sufficiently homogeneous (e.g. for self-similar fractal index sets or for T a finite-dimensional Euclidean ball with a Hölder-type canonical metric), but Dudley's bound can be strictly larger than γ_2 — and hence strictly larger than the true value of $\mathbb{E}[\sup_t X_t]$ — on highly inhomogeneous index sets, where the optimal admissible sequence of partitions in the chaining functional adapts the resolution of the chaining tree non-uniformly across T in a way that a single global entropy integral cannot capture. Talagrand's theorem (1987, with the matching two-sided bound established via the majorizing measure theorem) settled the long-standing question of whether *any* purely metric quantity could characterise $\mathbb{E}[\sup_t X_t]$ up to universal constants; γ_2 is precisely such a quantity, and no simpler one exists.

3.4 The Borell–TIS Inequality

Beyond the Dudley bound for the mean, we need tail bounds for the probability that $\sup_t X_t$ exceeds its mean by a large amount.

Theorem: Borell–Tsirelson–Ibragimov–Sudakov (Borell–TIS)

Let $(X_t)_{t\in T}$ be a mean-zero separable Gaussian field with $\sigma^2 = \sup_{t\in T} \operatorname{Var}(X_t) < \infty$ and $M = \mathbb{E}[\sup_t X_t] < \infty$. Then for all $u > 0$,

$$\mathbb{P}\left(\sup_{t\in T} X_t \geq M + u\right) \leq \exp\left(-\frac{u^2}{2\sigma^2}\right).$$

Proof sketch. Define $F : \mathbb{R}^n \to \mathbb{R}$ by $F(\mathbf{x}) = \sup_t \sum_i x_i e_i(t)$ where $\{e_i\}$ is an ONB of the Cameron–Martin space. Then F is σ-Lipschitz (in the Euclidean norm on coefficients) and $F(Z)$ has the same distribution as $\sup_t X_t$ when $Z \sim \mathcal{N}(0, I)$. The Gaussian isoperimetric inequality (Borell, 1975) gives the stated tail bound. □

Chapter 4

The Gaussian Free Field

4.1 Definition and Construction

The **Gaussian Free Field** (GFF) is the canonical infinite-dimensional Gaussian measure associated to the Laplacian, and it plays the role in two-dimensional quantum field theory that Brownian motion plays in one dimension.

Definition: Gaussian Free Field on a Domain

Let $D \subset \mathbb{R}^2$ be a bounded simply-connected domain with smooth boundary. The **Gaussian Free Field** Φ on D with Dirichlet boundary conditions is the Gaussian field formally defined by

$$d\mathbb{P}(\Phi) \propto \exp\left(-\frac{1}{2}\int_D |\nabla\phi(x)|^2\,dx\right)\prod_{x\in D} d\phi(x),$$

with covariance function equal to the Green's function $G_D(x,y)$ of $-\Delta$ on D:

$$\operatorname{Cov}(\Phi(x),\Phi(y)) = G_D(x,y).$$

The GFF is not a function but a distribution: $\Phi \notin L^2(D)$ almost surely. The precise construction proceeds via the eigenfunctions of $-\Delta$ on D.

Let (λ_n, e_n) be the eigenvalue-eigenfunction pairs of $-\Delta$ on D with Dirichlet boundary conditions, so $-\Delta e_n = \lambda_n e_n$, $\{e_n\}$ is an ONB for $L^2(D)$, and $\lambda_1 < \lambda_2 \le \lambda_3 \le \cdots \to \infty$.

Proposition: Series Representation of the GFF

Let $\{\xi_n\}_{n\ge 1}$ be i.i.d. $\mathcal{N}(0,1)$. The series

$$\Phi = \sum_{n=1}^{\infty} \frac{\xi_n}{\sqrt{\lambda_n}} e_n$$

converges in the Sobolev space $H^{-\varepsilon}(D)$ for all $\varepsilon > 0$, almost surely, and defines the GFF on D.

Remark: Logarithmic Covariance

On the unit disc $D = \mathbb{D}$, the Green's function is $G_D(x, y) = -\frac{1}{2\pi} \log |x - y| + h(x, y)$ where h is harmonic. Hence $\mathrm{Cov}(\Phi(x), \Phi(y)) \sim -\frac{1}{2\pi} \log |x - y|$ as $|x - y| \to 0$: the GFF has logarithmically divergent variance. This is analogous to Brownian motion's linear variance growth.

4.2 Conformal Invariance of the GFF

A key property that distinguishes the GFF in two dimensions is its invariance under conformal maps, making it central to conformal field theory and SLE (Schramm–Loewner Evolution).

Theorem: Conformal Covariance of the GFF

Let $\varphi : D \to D'$ be a conformal map between two bounded simply-connected planar domains. If Φ is the GFF on D with Dirichlet boundary conditions, then $\Phi \circ \varphi^{-1}$ is the GFF on D' in distribution. Equivalently,

$$G_D(x, y) = G_{D'}(\varphi(x), \varphi(y)).$$

Example: GFF on the Half-Plane

On the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$, the Green's function with Dirichlet boundary conditions along $\mathbb{R}$ is

$$G_{\mathbb{H}}(z, w) = \frac{1}{2\pi} \log \left| \frac{z - \bar{w}}{z - w} \right|.$$

Via the Riemann mapping theorem, every simply-connected domain $D \neq \mathbb{C}$ is conformally equivalent to $\mathbb{H}$, so the GFF on D is the pullback of the GFF on $\mathbb{H}$ under any conformal map $D \to \mathbb{H}$.

4.3 Thick Points, Level Sets, and Liouville Quantum Gravity

The GFF has remarkable geometric features: its level sets (zero-height contours) are conformally invariant curves related to SLE_4, and exponential functionals of the GFF define the Liouville quantum gravity measure.

Definition: Liouville Quantum Gravity Measure

For $\gamma \in (0, 2)$, the **Liouville quantum gravity (LQG) measure** on D is formally defined as

$$M_\gamma(dx) = e^{\gamma \Phi(x)}\, dx,$$

where dx is the Lebesgue measure on D. Since Φ is a distribution (not a function), this requires regularisation: one sets $\Phi_\varepsilon(x) = \mathbb{E}[\Phi(x) \mid \sigma(\Phi \text{ averaged over balls of radius } \varepsilon)]$ and defines

$$M_\gamma(dx) = \lim_{\varepsilon \to 0} \varepsilon^{\gamma^2/2} e^{\gamma \Phi_\varepsilon(x)}\, dx$$

where the limit holds in probability as a random Borel measure.

Theorem: Duplantier–Sheffield

For $\gamma \in (0, 2)$, the LQG measure M_γ is a non-trivial, non-atomic random Borel measure on D, almost surely of infinite total mass but locally finite. It is conformally covariant: if $\varphi : D \to D'$ is a conformal map, then M_γ pushes forward under φ to the LQG measure on D' up to a boundary correction term.

Part II

Point Processes

Chapter 5

Poisson Point Processes

5.1 Definition and Fundamental Properties

Point processes model the random placement of a countable collection of points in a space $\mathcal{X}$, and the Poisson process is the canonical model of completely random, non-interacting points.

Definition: Point Process

Let $(\mathcal{X}, \mathcal{B})$ be a locally compact, second-countable Hausdorff space with its Borel σ-algebra. A **point process** on $\mathcal{X}$ is a random variable Φ taking values in the space $\mathbf{N}(\mathcal{X})$ of locally finite counting measures on $\mathcal{X}$:

$$\Phi = \sum_i \delta_{X_i}, \qquad X_i \in \mathcal{X},$$

where $\Phi(B) = \#\{i : X_i \in B\} < \infty$ for all bounded $B \in \mathcal{B}$.

Definition: Poisson Point Process

A point process Φ on $\mathcal{X}$ is a **Poisson point process** (PPP) with **intensity measure** Λ (a locally finite, non-negative Borel measure on $\mathcal{X}$) if:

(i) **Distribution of counts**: for every bounded $B \in \mathcal{B}$, $\Phi(B) \sim \text{Poisson}(\Lambda(B))$;

(ii) **Independence**: for any finite collection of pairwise disjoint bounded Borel sets $B_1, \ldots, B_k$, the random variables $\Phi(B_1), \ldots, \Phi(B_k)$ are mutually independent.

We write $\Phi \sim \text{PPP}(\Lambda)$.

When $\mathcal{X} = \mathbb{R}^d$ and $\Lambda = \lambda \cdot \text{Leb}$ for some constant $\lambda > 0$ (the **intensity** or **rate**), Φ is called a **homogeneous** Poisson process with intensity λ.

Theorem: Existence and Uniqueness of the PPP

For every locally finite measure Λ on a locally compact second-countable space $\mathcal{X}$, there exists a Poisson point process $\Phi \sim \mathrm{PPP}(\Lambda)$, unique in distribution.

Proof. **Construction.** Write $\mathcal{X} = \bigcup_n B_n$ as a countable union of bounded Borel sets B_n with $\Lambda(B_n) < \infty$. On B_n, draw $N_n \sim \mathrm{Poisson}(\Lambda(B_n))$ points independently from the distribution $\Lambda(\cdot \cap B_n)/\Lambda(B_n)$. Amalgamate over n (the B_n may be taken disjoint). Independence across disjoint sets is built in by construction.

Uniqueness. The Laplace functional $L_\Phi(f) = \mathbb{E}[\exp(-\Phi(f))]$ for $f \geq 0$ measurable determines the law of Φ. For a PPP, $L_\Phi(f) = \exp(-\int (1 - e^{-f(x)}) d\Lambda(x))$, which is uniquely determined by Λ. □

5.2 The Laplace Functional and Void Probabilities

Definition: Laplace Functional

The **Laplace functional** of a point process Φ is

$$\mathcal{L}_\Phi[f] = \mathbb{E}\left[\exp\left(-\int_{\mathcal{X}} f(x)\, \Phi(dx)\right)\right] = \mathbb{E}\left[\exp\left(-\sum_i f(X_i)\right)\right],$$

defined for all measurable $f : \mathcal{X} \to [0, \infty)$.

Theorem: Laplace Functional of the PPP

For $\Phi \sim \mathrm{PPP}(\Lambda)$,

$$\mathcal{L}_\Phi[f] = \exp\left(-\int_{\mathcal{X}} \left(1 - e^{-f(x)}\right) d\Lambda(x)\right).$$

Proof. By independence of Φ on disjoint sets, on each small region dx,

$$\mathcal{L}_\Phi[f] = \prod_{dx} \mathbb{E}[e^{-f(x)\Phi(dx)}].$$

Since $\Phi(dx) \in \{0, 1\}$ for infinitesimally small dx (no double points a.s. when Λ is diffuse),

$$\mathbb{E}[e^{-f(x)\Phi(dx)}] = 1 - (1 - e^{-f(x)})\Lambda(dx) \approx e^{-(1-e^{-f(x)})\Lambda(dx)}.$$

Taking the product over all dx gives the result. □

Corollary: Void Probabilities

For $B \in \mathcal{B}$, the probability that Φ has no points in B is

$$\mathbb{P}(\Phi(B) = 0) = e^{-\Lambda(B)}.$$

5.3 The Campbell Theorem

Theorem: Campbell's Theorem

Let $\Phi \sim \text{PPP}(\Lambda)$ and $f : \mathcal{X} \to \mathbb{R}$ a measurable function. Then:

(i) **First moment**: $\mathbb{E}[\sum_i f(X_i)] = \int_{\mathcal{X}} f(x)\, d\Lambda(x)$;

(ii) **Second moment**: $\text{Var}(\sum_i f(X_i)) = \int_{\mathcal{X}} f(x)^2\, d\Lambda(x)$;

(iii) **Cumulant generating function**: for $\theta \in \mathbb{R}$ with $\int |e^{\theta f} - 1| d\Lambda < \infty$,

$$\log \mathbb{E}\left[e^{\theta \sum_i f(X_i)}\right] = \int_{\mathcal{X}} \left(e^{\theta f(x)} - 1\right) d\Lambda(x).$$

Example: Shot Noise Process

In wireless network analysis, the signal-to-interference ratio at a receiver located at the origin is

$$I = \sum_{X_i \in \Phi} \|X_i\|^{-\alpha} S_i,$$

where $\Phi \sim \text{PPP}(\lambda \cdot \text{Leb})$ is the set of interferer locations, $\alpha > 2$ is the path-loss exponent, and S_i are i.i.d. fading coefficients. Campbell's theorem gives $\mathbb{E}[I] = \lambda \int_{\mathbb{R}^2} \|x\|^{-\alpha}\, \mathbb{E}[S] dx$; this integral converges when $\alpha > 2$. The Laplace transform of I is $\mathcal{L}_I(s) = \exp(-\lambda \pi \Gamma(1 - 2/\alpha)(\mathbb{E}[S]s)^{2/\alpha})$ for Rayleigh fading, by direct computation via the Laplace functional.

5.4 Superposition, Thinning, and the Mapping Theorem

Theorem: Superposition

If $\Phi_1 \sim \text{PPP}(\Lambda_1)$ and $\Phi_2 \sim \text{PPP}(\Lambda_2)$ are independent, then $\Phi_1 + \Phi_2 \sim \text{PPP}(\Lambda_1 + \Lambda_2)$.

Theorem: Independent Thinning

Let $\Phi \sim \text{PPP}(\Lambda)$ and retain each point X_i independently with probability $p(X_i) \in [0, 1]$. The retained points form a $\text{PPP}(p \cdot \Lambda)$ and the deleted points form an independent $\text{PPP}((1 - p) \cdot \Lambda)$.

Theorem: Mapping Theorem

Let $\Phi \sim \text{PPP}(\Lambda)$ on $\mathcal{X}$ and $g : \mathcal{X} \to \mathcal{Y}$ a measurable function. Then $g(\Phi) = \sum_i \delta_{g(X_i)}$ is a $\text{PPP}(g_* \Lambda)$ on $\mathcal{Y}$, where $g_* \Lambda$ is the pushforward of Λ under g.

5.5 Palm Calculus

Palm calculus provides the conditional law of a point process given that a point exists at a specific location—an essential tool for second-order analysis of spatial patterns.

Definition: Palm Distribution

The **Palm distribution** $\mathbb{P}^x$ of a stationary point process Φ is the conditional law of Φ given that Φ has a point at x. More precisely, for the reduced Palm distribution $\mathbb{P}^{!x}$, one removes the point at x and conditions on the remaining configuration. For a stationary process, $\mathbb{P}^x$ does not depend on x by stationarity, and we write $\mathbb{P}^0$ for the Palm distribution at the origin.

Theorem: Campbell–Little–Mecke (CLM) Formula

For a stationary point process Φ with intensity λ and any non-negative measurable function $h : \mathcal{X} \times \mathbf{N}(\mathcal{X}) \to [0, \infty)$,

$$\mathbb{E}\left[\sum_{X \in \Phi} h(X, \Phi)\right] = \lambda \int_{\mathcal{X}} \mathbb{E}^{!x}[h(x, \Phi \cup \{x\})]\, dx,$$

where $\mathbb{E}^{!x}$ denotes expectation under the reduced Palm distribution at x.

Corollary: Palm Distribution of the Poisson Process

For $\Phi \sim \mathrm{PPP}(\Lambda)$, the Palm distribution $\mathbb{P}^{!x}$ coincides with the original distribution $\mathbb{P}$ of Φ. That is, Φ under $\mathbb{P}^{!x}$ has the same distribution as Φ under $\mathbb{P}$. This property is called **Slivnyak's theorem** and characterizes the PPP among stationary point processes.

Chapter 6

Cox Processes and Clustering

6.1 Definition and Examples

Cox processes (doubly stochastic Poisson processes) generalise the Poisson process by allowing the intensity itself to be random, modeling the clustering of events driven by an unobserved latent field.

Definition: Cox Process

Let Λ be a random locally finite measure on $\mathcal{X}$ (the **driving measure**). A **Cox process** (or **doubly stochastic Poisson process**) directed by Λ is a point process Φ such that, conditionally on $\Lambda = \mu$, $\Phi \sim \mathrm{PPP}(\mu)$.

Theorem: Laplace Functional of a Cox Process

If Φ is a Cox process directed by Λ, then

$$\mathcal{L}_\Phi[f] = \mathcal{L}_\Lambda\left[1 - e^{-f}\right] = \mathbb{E}\left[\exp\left(-\int_{\mathcal{X}} (1 - e^{-f(x)})\, d\Lambda(x)\right)\right].$$

Example: Log-Gaussian Cox Process

Let $\Lambda(dx) = e^{Z(x)}dx$ where $Z \sim \mathcal{GP}(m, k)$ is a Gaussian random field. The resulting Cox process is called a **log-Gaussian Cox process** (LGCP). Its intensity function is $\lambda(x) = \mathbb{E}[e^{Z(x)}] = e^{m(x)+k(x,x)/2}$, and its pair correlation function is $g(x, y) = e^{k(x,y)}$. LGCPs are widely used in spatial epidemiology (modelling disease incidence driven by an unobserved environmental field), ecology (clustering of trees or animals), and seismology (aftershock sequences).

6.2 Pair Correlation Functions and Second-Order Analysis

Definition: Intensity and Pair Correlation Function

For a stationary point process Φ with intensity $\lambda > 0$, the **pair correlation function** $g : [0, \infty) \to [0, \infty)$ is defined (when it exists) by

$$\mathbb{E}\left[\sum_{\substack{X,Y \in \Phi \\ X \neq Y}} h(\|X - Y\|)\right] = \lambda^2 \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g(\|x - y\|)\, h(\|x - y\|)\, dx\, dy$$

for all measurable $h \geq 0$. The quantity $\lambda^2 g(r)$ is the expected density of ordered pairs at distance r.

For a homogeneous PPP, $g \equiv 1$ (no spatial interaction). For a clustered process, $g(r) > 1$ for small r (nearby points are more likely than Poisson). For a repulsive (regular) process, $g(r) < 1$ for small r.

6.3 Ripley's K-Function and Summary Statistics

Definition: Ripley's K-Function

For a stationary isotropic point process with intensity λ, **Ripley's K-function** is

$$K(r) = \frac{1}{\lambda}\mathbb{E}[\text{number of further points within distance } r \text{ of a typical point}] = \int_0^r \lambda g(s) \omega_d s^{d-1}\, ds,$$

where ω_d is the surface area of the unit sphere in $\mathbb{R}^d$. For a PPP, $K(r) = \omega_d r^d / d = \text{Vol}(B(0, r))$.

Example: Estimation of K in Practice

Given an observed point pattern $\{x_1, \ldots, x_n\}$ in a bounded window $W \subset \mathbb{R}^2$, the edge-corrected estimator of $K(r)$ is

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_{i \neq j} \mathbf{1}(\|x_i - x_j\| \leq r)\, w_{ij}^{-1},$$

where $|W|$ is the area of W and w_{ij} is an edge-correction weight (e.g., the fraction of the circumference of a circle centered at x_i through x_j that lies inside W). Comparing $\hat{K}(r)$ with πr^2 (the Poisson benchmark) reveals clustering ($\hat{K} > \pi r^2$) or regularity ($\hat{K} < \pi r^2$).

Chapter 7

Gibbs Point Processes

7.1 Interaction Potentials and the Gibbs Distribution

Gibbs point processes model pairwise (and higher-order) interactions between points, generalising both the Poisson process (no interaction) and hard-core processes (infinite repulsion at short distances).

Definition: Gibbs Point Process in a Bounded Window

Let $W \subset \mathbb{R}^d$ be a bounded Borel set and $\Lambda_0 = \lambda \cdot \mathrm{Leb}$ a reference Poisson process. A **Gibbs point process** on W with pair potential $\varphi : [0, \infty) \to (-\infty, +\infty]$ and activity $z > 0$ is the point process with density with respect to $\mathrm{PPP}(z \cdot \mathrm{Leb}_W)$ given by

$$p(\mathbf{x}) = \frac{1}{Z_W} \exp\left(- \sum_{\{i,j\} \subset \mathbf{x}} \varphi(\|x_i - x_j\|) \right),$$

where $Z_W = \sum_{n \geq 0} \frac{z^n}{n!} \int_{W^n} e^{-\sum_{\{i,j\}} \varphi(\|x_i - x_j\|)} dx_1 \cdots dx_n$ is the **partition function**.

Example: Hard-Core Strauss Process

The **Strauss process** with parameter $\gamma \in [0, 1]$ and interaction range $R > 0$ has pair potential

$$\varphi(r) = -\log \gamma \cdot \mathbf{1}(r \leq R).$$

The density is $p(\mathbf{x}) \propto z^{|\mathbf{x}|} \gamma^{s(\mathbf{x})}$ where $|\mathbf{x}|$ is the number of points and $s(\mathbf{x}) = \#\{\{i,j\} : \|x_i - x_j\| \leq R\}$ is the number of R-close pairs. When $\gamma = 1$, this is a PPP. When $\gamma = 0$, no two points may be within distance R: the **hard-core process**. The Strauss process is repulsive for all $\gamma \in [0, 1)$: pairs at distance $\leq R$ are penalized by factor $\gamma < 1$.

7.2 The DLR Equations and Infinite-Volume Gibbs Measures

On unbounded domains, the correct formulation of Gibbs measures requires avoiding the partition function (which may diverge) and instead specifying conditional distributions for finite regions given the configuration outside.

Theorem: Dobrushin–Lanford–Ruelle (DLR) Characterization

A stationary point process Φ on $\mathbb{R}^d$ is a **Gibbs process** for the potential φ if and only if for every bounded $\Lambda \subset \mathbb{R}^d$, the conditional law of $\Phi \cap \Lambda$ given $\Phi \cap \Lambda^c = \xi_{\Lambda^c}$ is the Gibbs distribution on Λ with boundary condition ξ_{Λ^c}:

$$\mathbb{P}(\Phi \cap \Lambda \in \cdot \mid \Phi \cap \Lambda^c = \xi_{\Lambda^c}) = \mu_{\Lambda}^{\xi_{\Lambda^c}}(\cdot).$$

This family of consistent conditional distributions (the DLR equations) characterises the Gibbs process without reference to a partition function.

7.3 MCMC Simulation of Gibbs Point Processes

The intractable partition function Z_W means that Gibbs point processes cannot be simulated by direct inversion. Markov chain Monte Carlo (MCMC) methods produce samples by constructing an ergodic chain whose stationary distribution is the target Gibbs measure.

The Metropolis–Hastings Birth-Death Algorithm

The **spatial birth-death algorithm** (Preston, 1977; Ripley, 1977) operates on the space $\mathbf{N}(W)$ of finite configurations in a bounded window W. Starting from an arbitrary configuration $\mathbf{x}^{(0)}$, each step proposes either a birth (insertion of a new point) or a death (removal of an existing point), and accepts or rejects the proposal according to a Metropolis ratio.

Definition: Metropolis–Hastings Birth-Death Chain

At each step, given the current configuration $\mathbf{x}$:

1. With probability p_b, propose a **birth**: draw a candidate location $u \sim \mathrm{Uniform}(W)$. Accept the birth (set $\mathbf{x} \leftarrow \mathbf{x} \cup \{u\}$) with probability

$$\alpha_b(\mathbf{x}, u) = \min\left(1,\ \frac{z\,|W|\,e^{-\sum_{x \in \mathbf{x}} \varphi(\|u-x\|)}}{p_b/p_d \cdot (|\mathbf{x}|+1)}\right);$$

2. With probability $p_d = 1 - p_b$, propose a **death**: choose a point $x_i \in \mathbf{x}$ uniformly at random. Accept the death (set $\mathbf{x} \leftarrow \mathbf{x} \setminus \{x_i\}$) with probability $\alpha_d = 1/\alpha_b(\mathbf{x} \setminus \{x_i\}, x_i)$.

The resulting chain is reversible with respect to the Gibbs measure μ on W.

Verification of detailed balance. The detailed balance condition requires $\mu(\mathbf{x})\, q(\mathbf{x}, \mathbf{x}\cup\{u\})\, \alpha_b(\mathbf{x}, u) = \mu(\mathbf{x}\cup\{u\})\, q(\mathbf{x}\cup\{u\}, \mathbf{x})\, \alpha_d(\mathbf{x}\cup\{u\}, u)$. By definition of the Gibbs density,

$$\frac{\mu(\mathbf{x}\cup\{u\})}{\mu(\mathbf{x})} = \frac{z\, e^{-\sum_{x\in\mathbf{x}} \varphi(\|u-x\|)}}{|\mathbf{x}|+1},$$

where the factor $1/(|\mathbf{x}|+1)$ arises from the symmetrisation over ordered tuples. The proposal densities are $q(\mathbf{x}, \mathbf{x}\cup\{u\}) = p_b/|W|$ and $q(\mathbf{x}\cup\{u\}, \mathbf{x}) = p_d/(|\mathbf{x}|+1)$. Substituting into the detailed balance equation and using the definition of the Metropolis ratio α_b confirms the equality. □

Remark: Perfect Simulation and the Dominated CFTP Algorithm

For monotone Gibbs processes (where the density is monotone decreasing in the number of points, as for repulsive Strauss processes), **dominated coupling from the past** (Kendall–Møller, 2000) produces *exact* samples from μ without MCMC bias. The algorithm sandwiches the target process between dominating and dominated processes and runs the chain backwards in time until coalescence, guaranteeing that the output is an exact draw from μ.

7.4 Neyman–Scott Cluster Processes

Cluster processes model point patterns where points aggregate around unobserved parent locations. They are a special case of Cox processes where the driving measure is a superposition of kernels centred on a hidden PPP.

Definition: Neyman–Scott Process

Let $\Phi_p \sim \mathrm{PPP}(\kappa \cdot \mathrm{Leb})$ be a **parent process** with intensity $\kappa > 0$. Given Φ_p, for each parent $c \in \Phi_p$ independently generate a finite **offspring process** Φ_c with $N_c \sim \mathrm{Poisson}(\mu)$ points placed i.i.d. according to a **dispersal kernel** $h(\cdot - c)$ centered on c. The **Neyman–Scott process** is

$$\Phi = \bigcup_{c\in\Phi_p} \Phi_c,$$

the superposition of all offspring. The observable process Φ is a Cox process directed by $\Lambda(dx) = \kappa\mu \int h(x-c)\, dc$.

Proposition: Pair Correlation of the Neyman–Scott Process

The Neyman–Scott process Φ is stationary with intensity $\lambda = \kappa\mu$ and pair correlation function

$$g(r) = 1 + \frac{1}{\kappa\mu^2}\,(h * \tilde{h})(r),$$

where $\tilde{h}(x) = h(-x)$ and $*$ denotes convolution. The excess clustering $g(r) - 1$ reflects the intra-cluster distribution h; it is always positive, confirming aggregation.

Proof. By Campbell's theorem applied to the parent process, the intensity of Φ is $\lambda = \kappa\mu \int h(x)dx = \kappa\mu$. For the pair correlation, one computes the second factorial moment of Φ by conditioning on Φ_p. Two offspring points X, Y from the same parent c contribute an intra-cluster term; two points from different parents contribute the background Poisson term. The intra-cluster contribution to the second factorial moment density is

$$\rho^{(2)}_{\text{intra}}(x,y) = \kappa\,\mu(\mu-1)\int h(x-c)h(y-c)dc + \kappa\mu\int h(x-c)\delta(x-y)dc,$$

where the first term counts ordered pairs from the same cluster with distinct offspring. Taking the limit as $\mu \to \infty$ with $\kappa\mu = \lambda$ fixed (i.e. many small clusters) and dropping the diagonal gives $\rho^{(2)}(x,y) = \lambda^2 g(\|x-y\|)$ with g as stated. □

Example: Thomas Process

The **Thomas process** is the Neyman–Scott process with Gaussian dispersal kernel $h(x) = (2\pi\sigma^2)^{-d/2}e^{-\|x\|^2/(2\sigma^2)}$. In $\mathbb{R}^2$, the pair correlation function is

$$g(r) = 1 + \frac{1}{4\pi\kappa\sigma^2}e^{-r^2/(4\sigma^2)}.$$

This has a maximum at $r = 0$ and decays to 1 as $r \to \infty$ at a Gaussian rate controlled by σ. The Thomas process is widely used in ecology to model the dispersal of seeds or offspring from parent organisms: κ is the density of parent trees, μ the mean number of offspring, and σ the dispersal distance. The parameters (κ, μ, σ) can be estimated by minimum contrast fitting of the empirical pair correlation function to the theoretical formula above.

Part III

Spatial Gibbs Measures and Phase Transitions

Chapter 8

The Ising Model and Phase Transitions

8.1 Definition of the Ising Model

The Ising model is the canonical model of a ferromagnet on a lattice, and the study of its phase transition is one of the great achievements of twentieth-century mathematical physics.

Definition: Ising Model on a Finite Graph

Let $G = (V, E)$ be a finite graph. A **spin configuration** is $\sigma \in \{-1, +1\}^V$. The **Hamiltonian** at inverse temperature $\beta > 0$ and external field $h \in \mathbb{R}$ is

$$H(\sigma) = -\sum_{\{i,j\} \in E} \sigma_i \sigma_j - h \sum_{i \in V} \sigma_i .$$

The **Ising (Gibbs) measure** is

$$\mu_{\beta,h}(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z(\beta, h)}, \qquad Z(\beta, h) = \sum_{\sigma \in \{-1,+1\}^V} e^{-\beta H(\sigma)} .$$

On the integer lattice $\mathbb{Z}^d$, the Ising model is defined in infinite volume via the DLR equations, taking finite-volume Gibbs measures with $\pm$ boundary conditions and passing to the limit.

Definition: Infinite-Volume Ising Measure

The $+$ **infinite-volume Ising measure** $\mu^{+}_{\beta,h}$ on $\mathbb{Z}^d$ is the weak limit of the finite-volume Gibbs measures on $[-n, n]^d \cap \mathbb{Z}^d$ with all boundary spins set to $+1$. Similarly, $\mu^{-}_{\beta,h}$ is the limit with all boundary spins set to -1. The model exhibits **spontaneous magnetisation** if $\mu^{+}_{\beta,0}(\sigma_0 = +1) > \tfrac{1}{2}$, equivalently if $\mu^{+}_{\beta,0} \neq \mu^{-}_{\beta,0}$.

8.2 The Peierls Argument: Phase Transition in $d \geq 2$

Theorem: Peierls (1936) — Phase Transition in $\mathbb{Z}^2$

For the Ising model on $\mathbb{Z}^2$ with $h = 0$, there exists a critical inverse temperature $\beta_c > 0$ such that:

- For $\beta > \beta_c$: $\mu^+_{\beta,0} \neq \mu^-_{\beta,0}$ (two distinct Gibbs measures; spontaneous magnetisation);
- For $\beta < \beta_c$: there is a unique Gibbs measure and no spontaneous magnetisation.

The exact value is $\beta_c = \frac{1}{2}\log(1+\sqrt{2})$ (Onsager, 1944).

Proof of phase transition for $\beta > \beta_c$ large (Peierls argument). Consider the finite box $\Lambda_n = [-n, n]^2 \cap \mathbb{Z}^2$ with $+$ boundary conditions. A spin configuration σ is in bijection with a collection of *contours*: closed curves on the dual lattice $(\mathbb{Z} + \frac{1}{2})^2$ separating $+$ spins from $-$ spins. Each contour of length $|\gamma|$ contributes an energy penalty of $2\beta|\gamma|$ (each bond across the contour appears in the Hamiltonian with the wrong sign). Therefore the probability of a contour γ surrounding the origin satisfies

$$\mu^+_{\beta,0}(\gamma \text{ present}) \leq e^{-2\beta|\gamma|}.$$

The number of contours of length ℓ surrounding the origin is at most $3^\ell \cdot \ell$ (at most 3 choices at each step, times ℓ possible starting points on the contour). Summing over $\ell \geq 4$,

$$\mu^+_{\beta,0}(\sigma_0 = -1) \leq \sum_{\ell \geq 4} 3^\ell \ell\, e^{-2\beta\ell} < \frac{1}{2}$$

for β large enough (specifically, for $e^{2\beta} > 3$ suffices). Hence $\mu^+_{\beta,0}(\sigma_0 = +1) > \frac{1}{2}$, establishing spontaneous magnetisation. □

8.3 The Lee–Yang Circle Theorem

Theorem: Lee–Yang (1952)

For the ferromagnetic Ising model (coupling constants $J_{ij} \geq 0$) on any finite graph at any $\beta > 0$, the zeros of the partition function $Z(\beta, z)$ (viewed as a polynomial in the complex fugacity $z = e^{2\beta h}$) all lie on the unit circle $|z| = 1$ in $\mathbb{C}$.

Remark: Consequences of the Lee–Yang Theorem

The Lee–Yang theorem has profound consequences:

(i) The **free energy** $f(\beta, h) = \lim_{n\to\infty} \frac{1}{|V_n|} \log Z(\beta, h)$ is analytic in h for all real $h \neq 0$, for any β. Phase transitions (non-analyticity of the free energy) can

only occur at $h = 0$.

(ii) The **magnetisation** $m(\beta, h) = \frac{\partial f}{\partial h}$ is analytic in h for $h \neq 0$.

(iii) Extensions of the Lee–Yang method (Lieb–Sokal, 1981) establish analyticity for more general spin systems.

8.4 FKG Inequality and Correlation Inequalities

Theorem: Fortuin–Kasteleyn–Ginibre (FKG) Inequality

Let μ be the Ising Gibbs measure with ferromagnetic couplings ($J_{ij} \geq 0$) and $h \geq 0$. If $f, g : \{-1, +1\}^V \to \mathbb{R}$ are both **increasing** (i.e., $\sigma \leq \sigma'$ componentwise implies $f(\sigma) \leq f(\sigma')$), then

$$\mu(fg) \geq \mu(f)\,\mu(g).$$

Increasing functions are positively correlated under μ.

Proof. It suffices to verify the FKG lattice condition: for all σ, σ',

$$\mu(\sigma \vee \sigma')\,\mu(\sigma \wedge \sigma') \geq \mu(\sigma)\,\mu(\sigma'),$$

where $\vee$ and $\wedge$ are coordinatewise max and min. For the Ising measure, $\log \mu(\sigma) = \beta(\sum_{ij} \sigma_i \sigma_j + h \sum_i \sigma_i) - \log Z$, and the inequality reduces to the super-modularity of the Hamiltonian under ferromagnetic couplings. $\square$

Part IV

Determinantal and Permanental Point Processes

Chapter 9

Determinantal Point Processes

9.1 Definition and Characterisation

Determinantal point processes (DPPs) model repulsive spatial patterns: points exhibit a systematic avoidance of each other due to the determinantal structure of their joint intensities. They arise naturally in random matrix theory, quantum mechanics, and combinatorics.

Definition: Determinantal Point Process

A point process Φ on a locally compact space $\mathcal{X}$ with reference measure μ is a **determinantal point process** (DPP) with **kernel** $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ if, for all $n \geq 1$ and all bounded measurable $f : \mathcal{X}^n \to \mathbb{R}$,

$$\mathbb{E}\left[\sum_{\substack{x_1,\ldots,x_n \in \Phi \\ \text{distinct}}} f(x_1, \ldots, x_n)\right] = \int_{\mathcal{X}^n} f(x_1, \ldots, x_n)\, \det[K(x_i, x_j)]_{1 \leq i,j \leq n}\, d\mu^n.$$

The n-th order **correlation function** is $\rho_n(x_1, \ldots, x_n) = \det[K(x_i, x_j)]_{1 \leq i,j \leq n}$.

Theorem: Existence and Characterisation (Macchi–Soshnikov)

A DPP with kernel K (self-adjoint as an operator on $L^2(\mathcal{X}, \mu)$) exists if and only if all eigenvalues λ_n of K lie in $[0, 1]$. The process is well-defined and has the Janossy densities (conditional densities)

$$j_n(x_1, \ldots, x_n) = \det[(I - K)^{-1} K(x_i, x_j)]$$

up to normalisation. The PPP is recovered when $K(x, y) = \lambda \cdot \delta(x - y)$ with $\lambda \in (0, 1)$.

9.2 The Gaussian Unitary Ensemble and Eigenvalue Processes

The prototypical DPP arises as the distribution of eigenvalues of a random Hermitian matrix.

Definition: Gaussian Unitary Ensemble

The **Gaussian Unitary Ensemble** (GUE) of size n is the probability measure on $n \times n$ Hermitian matrices H with density

$$p(H) \propto \exp\Big(-\frac{n}{2} \operatorname{tr} H^2\Big).$$

Equivalently, $H_{jj} \sim \mathcal{N}(0, 1/n)$ and $H_{jk} = \overline{H_{kj}} \sim (\mathcal{N}(0, 1/(2n)) + i\mathcal{N}(0, 1/(2n)))/\sqrt{2}$ independently for $j < k$.

Theorem: Eigenvalue Distribution of GUE

The joint density of the ordered eigenvalues $\lambda_1 \le \cdots \le \lambda_n$ of a GUE matrix is

$$p(\lambda_1, \dots, \lambda_n) = C_n \prod_{i<j} (\lambda_j - \lambda_i)^2 \prod_{i=1}^{n} e^{-n\lambda_i^2/2},$$

where C_n is a normalisation constant. The factor $\prod_{i<j}(\lambda_j - \lambda_i)^2 = (\det[\lambda_i^{j-1}])^2$ is the square of the Vandermonde determinant.

Theorem: GUE Eigenvalues as a DPP

The eigenvalue point process of the GUE is a DPP on $\mathbb{R}$ with kernel

$$K_n(x, y) = \sum_{k=0}^{n-1} p_k(x) p_k(y) e^{-n(x^2+y^2)/4},$$

where p_k are the normalised Hermite polynomials. In the bulk (x, y near 0) as $n \to \infty$, the rescaled kernel converges to the **sine kernel**:

$$K_{\text{sine}}(x, y) = \frac{\sin(\pi(x - y))}{\pi(x - y)}.$$

9.3 Repulsiveness and the Pair Correlation Function

Proposition: Repulsive Pair Correlation of DPPs

For a DPP with kernel K, the pair correlation function satisfies

$$g(x,y) = 1 - \frac{|K(x,y)|^2}{K(x,x)K(y,y)} \leq 1.$$

Hence $g(x,x) = 0$: two points can never coincide (the DPP is **simple**, with no multiple points), and nearby points are less likely than Poisson (repulsion).

Example: Zeros of Gaussian Analytic Functions

Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ be a random analytic function on the unit disc $\mathbb{D}$, where $a_n \sim \mathcal{N}_{\mathbb{C}}(0,1)$ i.i.d. The zeros of f form a DPP on $\mathbb{D}$ with the Bergman kernel $K(z,w) = (1 - z\bar{w})^{-2}$ and the hyperbolic area measure. This process is invariant under Möbius transformations of the disc, a property that characterises it uniquely among isometry-invariant DPPs on $\mathbb{D}$.

9.4 Sampling from DPPs

Definition: L-ensemble

For a finite set $\mathcal{Y} = \{1, \dots, N\}$ and a positive semi-definite matrix $L \in \mathbb{R}^{N \times N}$, the **$L$-ensemble DPP** assigns to each subset $Y \subseteq \mathcal{Y}$ a probability

$$\mathbb{P}(Y) = \frac{\det L_Y}{\det(I + L)},$$

where $L_Y = (L_{ij})_{i,j \in Y}$ is the principal submatrix indexed by Y.

Theorem: Exact Sampling Algorithm for DPPs

Let $L = \sum_n \lambda_n v_n v_n^\top$ be the eigendecomposition of L. To sample $Y \sim \mathrm{DPP}(L)$:

1. For each n, independently include v_n in a set J with probability $\lambda_n/(1+\lambda_n)$;
2. Run a Gram–Schmidt orthogonalisation procedure to select $|J|$ items from $\{1, \dots, N\}$ with probabilities proportional to determinants of submatrices of $[v_n]_{n \in J}$.

This algorithm runs in $O(N|J|^2)$ time.

Part V

Gaussian Processes on Metric Spaces and Manifolds

Chapter 10

Gaussian Processes Beyond $\mathbb{R}^d$

10.1 Kernels on Metric Spaces

The theory of Gaussian processes extends naturally to arbitrary metric spaces, with the positive-definite condition on the kernel remaining the key requirement. The most important cases are Riemannian manifolds, graphs, and function spaces.

Definition: Positive-Definite Kernel on a Metric Space

Let (M, d) be a metric space. A function $k : M \times M \to \mathbb{R}$ is a **positive-definite kernel on** M if it is symmetric and for all $n \geq 1$, all $x_1, \ldots, x_n \in M$, and all $c_1, \ldots, c_n \in \mathbb{R}$, $\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0$.

Theorem: Schoenberg's Characterisation

A continuous function $f : [0, \infty) \to \mathbb{R}$ with $f(0) = 1$ is positive-definite (i.e., $k(x, y) = f(d(x, y))$ is a positive-definite kernel) on *every* metric space (M, d) if and only if f is a completely monotone function: $(-1)^n f^{(n)}(t) \geq 0$ for all $t > 0$ and all $n \geq 0$.

Completely monotone functions are represented (by Bernstein's theorem) as Laplace transforms of non-negative measures: $f(t) = \int_0^\infty e^{-ts} d\nu(s)$. Examples include $f(t) = e^{-t^\alpha}$ for $\alpha \in (0, 1]$ and $f(t) = (1 + t)^{-\beta}$ for $\beta > 0$.

10.2 Graph-Indexed Gaussian Processes and the Gaussian Markov Random Field

Definition: Gaussian Markov Random Field

A Gaussian random field $X = (X_v)_{v \in V}$ indexed by the vertices of a graph $G = (V, E)$ is a **Gaussian Markov Random Field** (GMRF) with precision matrix $Q > 0$ if $X \sim \mathcal{N}(0, Q^{-1})$ and the sparsity pattern of Q reflects the graph structure: $Q_{uv} = 0$ whenever $\{u, v\} \notin E$.

Theorem: Conditional Independence and the Graph

For a GMRF with precision matrix Q, $X_u \perp\!\!\!\perp X_v \mid (X_w)_{w \neq u,v}$ if and only if $Q_{uv} = 0$, i.e., $\{u, v\} \notin E$.

The proof follows from the formula for conditional Gaussian distributions: conditioning on all other variables results in a bivariate Gaussian for (X_u, X_v) with conditional covariance determined by Q^{-1} entry, but the partial correlation is $-Q_{uv}/\sqrt{Q_{uu}Q_{vv}}$, which vanishes iff $Q_{uv} = 0$.

Example: GMRF on a Lattice and the SPDE Approach

On the d-dimensional integer lattice $\mathbb{Z}^d$ with $V = [-n, n]^d \cap \mathbb{Z}^d$ and $Q = \kappa^2 I - \Delta_G$ where Δ_G is the graph Laplacian, the resulting GMRF approximates the Matérn-1 Gaussian field. More precisely (Lindgren–Rue–Lindström, 2011), the solution to the SPDE

$$(\kappa^2 - \Delta)^{\nu/2} u = \mathcal{W} \qquad \text{on } \mathbb{R}^d,$$

where $\mathcal{W}$ is a spatial white noise, is a Gaussian field with Matérn-ν covariance. Discretising the SPDE on a mesh gives a sparse GMRF that can be inverted in $O(n^{d/2})$ time using sparse Cholesky factorisation, enabling scalable inference for large spatial datasets.

10.3 Gaussian Processes on Riemannian Manifolds

Definition: Heat Kernel and GP on a Manifold

Let (M, g) be a compact Riemannian manifold with Laplace–Beltrami operator Δ_g and eigenvalues $0 = \mu_0 < \mu_1 \leq \mu_2 \leq \cdots$ with corresponding orthonormal eigenfunctions $\phi_n \in L^2(M, \mathrm{vol}_g)$. The **heat kernel** is

$$p_t(x, y) = \sum_{n=0}^{\infty} e^{-\mu_n t}\, \phi_n(x)\, \phi_n(y),$$

and the **heat kernel Gaussian process** with time parameter $t > 0$ has covariance $k(x, y) = p_t(x, y)$.

Proposition: Matérn Kernels on Manifolds via the Spectral Approach

The Matérn-ν kernel on a compact Riemannian manifold (M, g) is defined spectrally by

$$k_\nu(x, y) = \sum_{n=0}^{\infty} \left(\frac{2\nu}{\ell^2} + \mu_n \right)^{-(\nu + d/2)} \phi_n(x) \phi_n(y),$$

which generalises the Euclidean Matérn covariance to arbitrary manifolds. This is the covariance of the solution to the manifold SPDE $(\kappa^2 - \Delta_g)^{(\nu+d/2)/2} u = \mathcal{W}_g$, where $\mathcal{W}_g$ is white noise with respect to vol_g.

Example: GP Regression on the Sphere $\mathbb{S}^2$

On the unit sphere $\mathbb{S}^2$, the eigenfunctions of $-\Delta_g$ are the spherical harmonics Y_ℓ^m with eigenvalues $\ell(\ell+1)$. A stationary Gaussian process on $\mathbb{S}^2$ (invariant under $SO(3)$) has covariance $k(x,y) = \sum_{\ell=0}^{\infty} h_\ell P_\ell(\cos\theta)$ where $\theta = d_{\mathbb{S}^2}(x,y)$ is the geodesic distance, P_ℓ are Legendre polynomials, and $h_\ell \geq 0$ are spectral coefficients. For the Matérn-ν kernel, $h_\ell \propto (\kappa^2 + \ell(\ell+1))^{-\nu-1}$. This representation enables efficient inference via the discrete spherical harmonic transform.

Part VI

Spatial Statistics and Bayesian Nonparametrics

Chapter 11

Kriging and Spatial Interpolation

11.1 The Kriging Predictor

Kriging (named after the South African mining engineer D.G. Krige) is the method of optimal linear unbiased prediction for spatial data under a Gaussian field model.

Definition: Simple Kriging

Suppose $X \sim \mathcal{GP}(0, k)$ is observed at locations $\mathbf{s} = (s_1, \dots, s_n)^\top$ with observations $\mathbf{y} = (X_{s_1}, \dots, X_{s_n})^\top$. The **simple kriging** predictor of X_{s^*} at a new location s^* is

$$\hat{X}_{s^*} = k(s^*, \mathbf{s})^\top K^{-1} \mathbf{y},$$

where $k(s^*, \mathbf{s}) = (k(s^*, s_1), \dots, k(s^*, s_n))^\top$ and $K_{ij} = k(s_i, s_j)$. The kriging variance is

$$\sigma^2_{s^*} = k(s^*, s^*) - k(s^*, \mathbf{s})^\top K^{-1} k(s^*, \mathbf{s}).$$

Theorem: Optimality of Kriging

Among all linear predictors $\hat{X}_{s^*} = \mathbf{a}^\top \mathbf{y}$ of X_{s^*}, the simple kriging predictor minimises the mean squared prediction error $\mathbb{E}[(X_{s^*} - \hat{X}_{s^*})^2]$. The minimum MSPE equals $\sigma^2_{s^*}$.

Proof. For an arbitrary linear predictor $\mathbf{a}^\top \mathbf{y}$,

$$\mathbb{E}[(X_{s^*} - \mathbf{a}^\top \mathbf{y})^2] = k(s^*, s^*) - 2\mathbf{a}^\top k(s^*, \mathbf{s}) + \mathbf{a}^\top K \mathbf{a}.$$

Minimising over $\mathbf{a}$ by taking the gradient and setting to zero: $K\mathbf{a} = k(s^*, \mathbf{s})$, so $\mathbf{a}^* = K^{-1} k(s^*, \mathbf{s})$. Substituting gives the kriging variance. $\square$

11.2 Universal Kriging and the Intrinsic Model

When the mean function $m(s) = \mathbb{E}[X_s]$ is not zero but belongs to a parametric family $m(s) = \mathbf{f}(s)^\top \boldsymbol{\beta}$ (e.g. polynomial in the coordinates), the predictor must account for the unknown $\boldsymbol{\beta}$.

Theorem: Universal Kriging (BLUP)

Let $X_s = \mathbf{f}(s)^\top \boldsymbol{\beta} + Z_s$ where $\boldsymbol{\beta}$ is unknown and $Z \sim \mathcal{GP}(0, k)$. The **best linear unbiased predictor** (BLUP) of X_{s^*} is

$$\hat{X}^{\mathrm{UK}}_{s^*} = \mathbf{f}(s^*)^\top \hat{\boldsymbol{\beta}} + k(s^*, \mathbf{s})^\top K^{-1}(\mathbf{y} - F\hat{\boldsymbol{\beta}}),$$

where $F_{ij} = f_j(s_i)$ is the regression matrix and $\hat{\boldsymbol{\beta}} = (F^\top K^{-1} F)^{-1} F^\top K^{-1}\mathbf{y}$ is the generalised least squares estimator.

11.3 Covariance Parameter Estimation by Maximum Likelihood

In practice, the covariance kernel $k(\cdot,\cdot;\boldsymbol{\theta})$ depends on unknown parameters $\boldsymbol{\theta}$ (e.g., the variance σ^2 and length-scale ℓ of a Matérn kernel). These are estimated by maximum likelihood.

Definition: Restricted Maximum Likelihood (REML)

Given observations $\mathbf{y} \sim \mathcal{N}(F\boldsymbol{\beta}, K(\boldsymbol{\theta}))$, the **restricted log-likelihood** is

$$\ell_{\mathrm{REML}}(\boldsymbol{\theta}) = -\frac{1}{2}\log\det K - \frac{1}{2}\log\det(F^\top K^{-1} F) - \frac{1}{2}\mathbf{y}^\top P \mathbf{y},$$

where $P = K^{-1} - K^{-1}F(F^\top K^{-1}F)^{-1}F^\top K^{-1}$ is the residual-maker. Maximising ℓ_{REML} over $\boldsymbol{\theta}$ gives the REML estimator $\hat{\boldsymbol{\theta}}_{\mathrm{REML}}$, which accounts for the estimation of $\boldsymbol{\beta}$ and typically has less bias than pure MLE for the variance components.

Chapter 12

Gaussian Process Regression and Bayesian Nonparametrics

12.1 Bayesian Inference with Gaussian Process Priors

Gaussian process regression (GPR) provides a principled Bayesian nonparametric approach to function estimation: rather than specifying a finite-dimensional parametric model for an unknown function f, we place a GP prior over the function space and update it using data.

Definition: Gaussian Process Regression Model

The **Gaussian process regression** (GPR) model for observations $y_i = f(x_i) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ i.i.d. noise is:

$$f \sim \mathcal{GP}(m, k) \quad \text{(prior)},$$
$$y_i \mid f \sim \mathcal{N}(f(x_i), \sigma_n^2) \quad \text{(likelihood)}.$$

Theorem: Posterior GP Distribution

Given observations $\mathbf{y} = (y_1, \dots, y_n)^\top$ at inputs $\mathbf{x} = (x_1, \dots, x_n)^\top$, the posterior distribution of f is again a Gaussian process:

$$f \mid \mathbf{y} \sim \mathcal{GP}(m_{\text{post}}, k_{\text{post}}),$$

with:

$$m_{\text{post}}(x^*) = m(x^*) + k(x^*, \mathbf{x})(K + \sigma_n^2 I)^{-1}(\mathbf{y} - m(\mathbf{x})),$$
$$k_{\text{post}}(x^*, x^{**}) = k(x^*, x^{**}) - k(x^*, \mathbf{x})(K + \sigma_n^2 I)^{-1} k(\mathbf{x}, x^{**}).$$

Proof. By conjugacy of Gaussian distributions: the joint distribution of $(f(\mathbf{x}), f(x^*))$ is Gaussian, and conditioning on $\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\varepsilon}$ is equivalent to conditioning on a Gaussian linear transformation of the joint, which preserves Gaussianity. The posterior mean and covariance follow from the standard conditional Gaussian formula (Proposition 4.1.1 of this volume). $\square$

12.2 Convergence of GP Regression: Posterior Contraction

A fundamental question in Bayesian nonparametrics is whether the posterior distribution concentrates around the true function as more data are observed.

Theorem: Posterior Contraction for GP Regression (Van der Vaart–Van Zanten)

Let $f_0 \in \mathcal{H}_{k_\nu}$, the RKHS of the Matérn-ν kernel on $[0,1]^d$, with $\nu > d/2$. Suppose n i.i.d. observations $y_i = f_0(x_i) + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are given, and the GP prior has Matérn-ν kernel with length-scale $\ell_n \asymp n^{-1/(2\nu+d)}$. Then the posterior contracts at the minimax optimal rate:

$$\mathbb{P}^n_{f_0}\left(\| f - f_0 \|_{L^2} > M_n \, n^{-\nu/(2\nu+d)} \right) \to 0$$

for any $M_n \to \infty$, where $\mathbb{P}^n_{f_0}$ is the n-fold product of the data distribution under f_0.

Remark: Optimal Rates and Adaptation

The rate $n^{-\nu/(2\nu+d)}$ matches the minimax rate for estimating Sobolev-ν functions on $[0,1]^d$, confirming that GP regression with the correct Matérn kernel is rate-optimal. An important open problem is **adaptation**: constructing a GP prior that achieves near-optimal rates without knowing ν in advance. Hierarchical priors over ν (or equivalently over the length-scale) are studied in this context.

12.3 Scalable Gaussian Process Methods

Exact GP regression requires $O(n^3)$ time for matrix inversion, which is prohibitive for $n > 10^4$. Several approximation strategies reduce this cost.

Definition: Sparse GP Approximations

Let $\mathbf{z} = (z_1, \ldots, z_m)^\top$ be a set of $m \ll n$ **inducing points** in the input space. The **Nyström approximation** replaces the kernel matrix $K \in \mathbb{R}^{n \times n}$ with the rank-m approximation

$$\tilde{K} = K_{\mathbf{xz}} K_{\mathbf{zz}}^{-1} K_{\mathbf{zx}},$$

where $K_{\mathbf{xz}} = (k(x_i, z_j))$ and $K_{\mathbf{zz}} = (k(z_i, z_j))$. The resulting approximate posterior is computed in $O(nm^2)$ time.

Theorem: Structured Kernel Interpolation (SKI)

If the inducing points $\mathbf{z}$ form a regular grid in $[0,1]^d$ with $m = m_1 \times \cdots \times m_d$ points, then the Nyström matrix $K_{\mathbf{xz}} K_{\mathbf{zz}}^{-1} K_{\mathbf{zx}}$ can be computed in $O(n + m \log m)$

time using the fast Fourier transform when the kernel is stationary, enabling GP regression at scale $n \sim 10^6$.

12.4 Variational Inference for Gaussian Processes

When the likelihood $p(\mathbf{y} \mid \mathbf{f})$ is non-Gaussian (e.g., for classification or count data), the GP posterior $p(\mathbf{f} \mid \mathbf{y})$ is no longer Gaussian and cannot be computed in closed form. Variational inference (VI) approximates the posterior by a tractable distribution chosen to minimise the Kullback–Leibler divergence.

Definition: Sparse Variational GP (SVGP)

Let $\mathbf{u} = f(\mathbf{z})$ be the GP values at m inducing points $\mathbf{z}$. The **sparse variational GP** (Titsias, 2009) approximates the posterior $p(\mathbf{f}, \mathbf{u} \mid \mathbf{y})$ by the variational family

$$q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u})\, q(\mathbf{u}), \qquad q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S}),$$

where $\mathbf{m} \in \mathbb{R}^m$ and $\mathbf{S} \succ 0$ are variational parameters, and $p(\mathbf{f} \mid \mathbf{u})$ is the exact GP conditional.

Theorem: Evidence Lower Bound for SVGP

The evidence lower bound (ELBO) for the SVGP model is

$$\mathcal{L}(\mathbf{m}, \mathbf{S}, \mathbf{z}) = \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{f})}[\log p(y_i \mid f_i)] - \mathrm{KL}(q(\mathbf{u}) \| p(\mathbf{u})),$$

where $q(f_i) = \mathcal{N}(\mu_i, \sigma_i^2)$ with

$$\mu_i = k(x_i, \mathbf{z}) K_{\mathbf{zz}}^{-1} \mathbf{m}, \quad \sigma_i^2 = k(x_i, x_i) - k(x_i, \mathbf{z}) K_{\mathbf{zz}}^{-1} k(\mathbf{z}, x_i) + k(x_i, \mathbf{z}) K_{\mathbf{zz}}^{-1} \mathbf{S}\, K_{\mathbf{zz}}^{-1} k(\mathbf{z}, x_i).$$

The KL divergence between two Gaussians is

$$\mathrm{KL}(q(\mathbf{u}) \| p(\mathbf{u})) = \tfrac{1}{2}\left[\mathrm{tr}(K_{\mathbf{zz}}^{-1}\mathbf{S}) + \mathbf{m}^\top K_{\mathbf{zz}}^{-1}\mathbf{m} - m + \log \frac{\det K_{\mathbf{zz}}}{\det \mathbf{S}} \right].$$

Jointly maximising $\mathcal{L}$ over $(\mathbf{m}, \mathbf{S}, \mathbf{z})$ and the kernel hyperparameters $\boldsymbol{\theta}$ by gradient ascent gives a scalable $O(nm^2)$ algorithm.

Derivation. By Jensen's inequality, $\log p(\mathbf{y}) = \log \int p(\mathbf{y}, \mathbf{f}, \mathbf{u}) d\mathbf{f}\, d\mathbf{u} \geq \mathbb{E}_{q(\mathbf{f},\mathbf{u})}[\log p(\mathbf{y}, \mathbf{f}, \mathbf{u}) - \log q(\mathbf{f}, \mathbf{u})]$. Substituting $p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{u}) p(\mathbf{u})$ and $q(\mathbf{f}, \mathbf{u}) = p(\mathbf{f} \mid \mathbf{u}) q(\mathbf{u})$, the $p(\mathbf{f} \mid \mathbf{u})$ terms cancel and one is left with $\mathcal{L} = \mathbb{E}_{q(\mathbf{f})}[\log p(\mathbf{y} \mid \mathbf{f})] - \mathrm{KL}(q(\mathbf{u}) \| p(\mathbf{u}))$. The marginal $q(f_i) = \int p(f_i \mid \mathbf{u}) q(\mathbf{u}) d\mathbf{u}$ is a Gaussian with the stated mean and variance by standard GP conditional formulas. $\square$

Example: GP Classification via Laplace Approximation

For binary classification with $y_i \in \{0, 1\}$ and likelihood $p(y_i \mid f_i) = \sigma(f_i)^{y_i}(1 - \sigma(f_i))^{1-y_i}$ where $\sigma(t) = 1/(1 + e^{-t})$ is the logistic function, the **Laplace approximation** to the GP posterior finds the MAP estimate $\mathbf{f}^* = \arg\min[-\log p(\mathbf{y} \mid \mathbf{f}) + \frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f}]$ (the mode of the unnormalised posterior) by Newton's method, then approximates the posterior as Gaussian centered at $\mathbf{f}^*$ with covariance $(K^{-1} + W)^{-1}$ where $W = -\nabla^2 \log p(\mathbf{y} \mid \mathbf{f}^*)$ is the negative Hessian of the log-likelihood. For logistic regression, $W = \operatorname{diag}(\sigma(f_i^*)(1 - \sigma(f_i^*)))$ is the diagonal matrix of Bernoulli variances evaluated at the mode.

Remark: Expectation Propagation

The Laplace approximation fits a Gaussian to the joint posterior at its mode. A more accurate alternative is **expectation propagation** (EP, Minka, 2001), which fits a Gaussian by matching moments factor-by-factor: each likelihood term $p(y_i \mid f_i)$ is approximated by a Gaussian "site" $\tilde{t}_i(f_i) = \tilde{z}_i \mathcal{N}(f_i; \tilde{\mu}_i, \tilde{\sigma}_i^2)$, and the site parameters are iteratively updated to match the marginal moments of the approximate posterior. EP typically gives more accurate predictive distributions than the Laplace approximation at the cost of additional inner iterations.

Chapter 13

Spatial Statistics: Testing and Model Selection

13.1 Testing for Complete Spatial Randomness

A fundamental task in spatial statistics is testing whether an observed point pattern is consistent with a homogeneous Poisson process (complete spatial randomness, CSR).

Definition: Envelope Test

Given an observed point pattern $\mathbf{x}$ in a window W, and a summary statistic $T(\mathbf{x})$ (such as $K(r)$ or $g(r)$), the **simulation envelope test** proceeds as follows:

1. Simulate B realisations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(B)}$ from the null model (CSR with intensity $\hat{\lambda} = n/|W|$);
2. Compute $T^{(b)} = T(\mathbf{x}^{(b)})$ for each simulation;
3. Reject the null if $T(\mathbf{x})$ lies outside the pointwise $(1-\alpha)$ simulation envelope $[\min_b T^{(b)}(r), \max_b T^{(b)}(r)]$ for some r.

Remark: Multiple Testing Correction

The envelope test as described above has inflated type-I error when evaluated at multiple values of r simultaneously. Corrections include the **global envelope test** (Myllymäki et al., 2017), which constructs simultaneous envelopes with exact significance level α by ranking curves rather than pointwise values.

13.2 Variogram Estimation and Model Fitting

Definition: Variogram

For an intrinsically stationary random field $\{X_s : s \in \mathbb{R}^d\}$ (where $\mathbb{E}[X_s - X_t]^2$ depends only on $s - t$), the **variogram** is

$$2\gamma(h) = \mathbb{E}\big[(X_{s+h} - X_s)^2\big], \qquad h \in \mathbb{R}^d.$$

The function $\gamma(h)$ (half the variogram) is called the **semivariogram**. When X is stationary with covariance C, $\gamma(h) = C(0) - C(h)$.

Proposition: Estimator of the Variogram

Given observations $\{x_{s_i}, i = 1, \dots, n\}$ (noiseless), the **method-of-moments variogram estimator** is

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} (x_{s_i} - x_{s_j})^2,$$

where $N(h) = \{(i, j) : s_i - s_j \approx h\}$ is the set of pairs at lag h (within a tolerance). This estimator is unbiased for $\gamma(h)$ and consistent as the number of observation pairs grows.

13.3 Spatial Prediction Under Model Uncertainty

Definition: Bayesian Model Averaging for Spatial Models

Given M competing covariance models $\{k_m : m = 1, \dots, M\}$ (e.g., Matérn with $\nu \in \{1/2, 3/2, 5/2, \infty\}$), the **Bayesian model average** (BMA) predictor at a new location s^* is

$$\hat{X}^{\text{BMA}}_{s^*} = \sum_{m=1}^{M} w_m \hat{X}^{(m)}_{s^*},$$

where $\hat{X}^{(m)}_{s^*}$ is the kriging predictor under model m and the model weights are

$$w_m = \frac{p(m)\, p(\mathbf{y} \mid m)}{\sum_{m'} p(m')\, p(\mathbf{y} \mid m')}$$

with $p(\mathbf{y} \mid m) = \int p(\mathbf{y} \mid \boldsymbol{\theta}, m) p(\boldsymbol{\theta} \mid m) d\boldsymbol{\theta}$ the marginal likelihood.

Part VII

Applications and Advanced Topics

Chapter 14

Applications in Physics, Biology, and Machine Learning

14.1 Statistical Physics: Random Fields as Field Theories

Definition: ϕ^4 Euclidean Field Theory

The ϕ^4 **Euclidean field theory** on $\mathbb{R}^d$ is formally defined by the measure

$$d\mu(\phi) \propto \exp\left(-\int_{\mathbb{R}^d} \left[\frac{1}{2}|\nabla\phi(x)|^2 + \frac{m^2}{2}\phi(x)^2 + \frac{\lambda}{4}\phi(x)^4 \right] dx \right) \mathcal{D}\phi,$$

where $\lambda > 0$ is the coupling constant and $m > 0$ the mass. When $\lambda = 0$, this reduces to the Gaussian measure with covariance $(-\Delta + m^2)^{-1}$, the massive GFF. For $\lambda > 0$, the ϕ^4 measure is a non-Gaussian random field that arises in the scaling limit of the Ising model above its upper critical dimension $d \geq 4$. Constructive quantum field theory (Glimm–Jaffe) establishes existence of the ϕ^4 measure in $d = 2, 3$ via correlation inequalities and cluster expansions.

14.2 Neuroscience and the Neural Field Model

Definition: Neural Field Equation

A **neural field** $u(x,t)$ (the mean firing rate of a population of neurons at position $x \in \mathbb{R}^d$ and time t) satisfies the Wilson–Cowan integro-differential equation

$$\tau \frac{\partial u}{\partial t}(x,t) = -u(x,t) + \int_{\mathbb{R}^d} w(x-y) f(u(y,t))\, dy + I(x,t),$$

where w is the spatial coupling kernel, f is the activation function, and I is the external input. When I is a random field and f is linear, the steady-state of u is itself a random field with covariance determined by w and the covariance of I

via a Wiener–Hopf integral equation.

14.3 Machine Learning: Deep Kernel Learning and Neural Networks as GPs

A major discovery in the theory of neural networks is that deep fully-connected networks converge to Gaussian processes in the infinite-width limit.

Theorem: Neural Network Gaussian Process (NNGP) Correspondence

Let $f^L : \mathbb{R}^d \to \mathbb{R}$ be a depth-L fully-connected neural network with i.i.d. random weights $W_{ij}^l \sim \mathcal{N}(0, \sigma_w^2/n_l)$ and biases $b_i^l \sim \mathcal{N}(0, \sigma_b^2)$, where n_l is the width of layer l. As $n_1, \dots, n_{L-1} \to \infty$ sequentially, f^L converges in distribution to a Gaussian process $\mathcal{GP}(0, k^L)$, where the kernel k^L is computed recursively:

$$k^1(x, x') = \sigma_w^2 x \cdot x'/d + \sigma_b^2,$$
$$k^{l+1}(x, x') = \sigma_w^2 \mathbb{E}_{(u,v)\sim\mathcal{N}(0,K^l)}[\phi(u)\phi(v)] + \sigma_b^2,$$

where $K^l = \begin{pmatrix} k^l(x,x) & k^l(x,x') \\ k^l(x',x) & k^l(x',x') \end{pmatrix}$ and ϕ is the pointwise activation function.

Example: ReLU NNGP Kernel

For the ReLU activation $\phi(t) = \max(0, t)$, the arc-cosine kernel at depth $L = 1$ is

$$k^2(x, x') = \frac{\sigma_w^2}{2\pi} \|x\| \, \|x'\| \, (\sin\theta + (\pi - \theta)\cos\theta) + \sigma_b^2,$$

where $\theta = \arccos(x \cdot x'/(\|x\| \, \|x'\|))$. This kernel captures angular similarity between inputs, reflecting the activation geometry of ReLU networks. Deeper compositions of this kernel are the NNGP kernels of deep ReLU networks.

Chapter 15

Stochastic Partial Differential Equations and Random Fields

15.1 White Noise and the Sobolev Scale

Definition: Spatial White Noise

Spatial white noise $\mathcal{W}$ on $\mathbb{R}^d$ is the centred Gaussian field indexed by $L^2(\mathbb{R}^d)$:

$$\mathcal{W}(f) = \int_{\mathbb{R}^d} f(x)\,\mathcal{W}(dx), \qquad \mathbb{E}[\mathcal{W}(f)\mathcal{W}(g)] = \langle f, g\rangle_{L^2(\mathbb{R}^d)}.$$

Formally, $\mathcal{W}$ has covariance "kernel" $\delta(x-y)$, but it is a distribution (Schwartz) rather than a function.

Theorem: SPDE Characterisation of Matérn Fields

The solution u of the SPDE

$$(\kappa^2 - \Delta)^{s/2} u = \mathcal{W} \qquad \text{on } \mathbb{R}^d, \quad s = \nu + d/2,$$

exists as an element of the Sobolev space $H^{s-d/2-\varepsilon}(\mathbb{R}^d)$ for all $\varepsilon > 0$, and its covariance is the Matérn-ν kernel:

$$\operatorname{Cov}(u(x), u(y)) = k_\nu(\|x-y\|).$$

This connection (Whittle, 1954; Lindgren–Rue–Lindström, 2011) is the basis for the SPDE approach to spatial statistics.

15.2 Stochastic Reaction-Diffusion Equations

Definition: Stochastic Heat Equation

The **stochastic heat equation** on $[0,1]^d$ with additive space-time white noise is

$$\frac{\partial u}{\partial t} = \Delta u + \dot{\mathcal{W}},$$

where $\dot{\mathcal{W}}(x,t)$ is space-time white noise: $\mathbb{E}[\dot{\mathcal{W}}(x,t)\dot{\mathcal{W}}(y,s)] = \delta(x-y)\delta(t-s)$. A rigorous formulation uses the mild solution: for $t > 0$,

$$u(x,t) = \int_0^1 p_t(x,y)u_0(y)\,dy + \int_0^t \int_{[0,1]^d} p_{t-s}(x,y)\,\mathcal{W}(dy\,ds),$$

where p_t is the heat kernel on $[0,1]^d$.

Theorem: Regularity of the Stochastic Heat Equation

For $d = 1$, the mild solution u of the stochastic heat equation exists as a continuous function, and the covariance is

$$\operatorname{Cov}(u(x,t),u(y,s)) = \int_0^{t\wedge s}\int_0^1 p_{t-r}(x,z)p_{s-r}(y,z)\,dz\,dr.$$

For $d \geq 2$, the mild solution does not exist as a function but only as a Schwartz distribution, since space-time white noise is too rough.

Chapter 16

Worked Applications

This chapter develops five self-contained applications of the theory of random fields and point processes, each carried through to explicit quantitative conclusions. The problems are chosen to cross Part boundaries: each draws simultaneously on kernels, GP inference, point process theory, and the spatial statistics machinery developed in the preceding chapters.

16.1 Kriging a Matérn Field: End-to-End Spatial Prediction

Setup

Let $\{Z(s) : s \in [0,1]^2\}$ be a zero-mean stationary Gaussian field with **Matérn-**$3/2$ covariance

$$k(s,t) = \sigma^2\left(1 + \frac{\sqrt{3}\,\|s-t\|}{\ell}\right)\exp\left(-\frac{\sqrt{3}\,\|s-t\|}{\ell}\right),$$

with variance $\sigma^2 > 0$ and length-scale $\ell > 0$. Observations $y_i = Z(s_i) + \varepsilon_i$ are available at n locations $s_1, \dots, s_n \in [0,1]^2$, with i.i.d. noise $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$.

Step 1: Explicit RKHS of the Matérn-$3/2$ kernel

By the SPDE characterisation (Chapter 15), the Matérn-$3/2$ field on $\mathbb{R}^2$ is the solution to

$$\left(\frac{2}{\ell^2} - \Delta\right) u = \mathcal{W},$$

(the case $\nu = 3/2$ corresponds to $s = \nu + d/2 = 2$, a second-order SPDE). Its RKHS $\mathcal{H}_k$ is the Sobolev space $H^2([0,1]^2)$, and the RKHS norm satisfies

$$\|f\|_{\mathcal{H}_k}^2 = \sigma^{-2}\int_{[0,1]^2}\left[\ell^4(\Delta f)^2 + 2\ell^2|\nabla f|^2 + f^2\right]dx.$$

Functions in $\mathcal{H}_k$ are continuously differentiable (C^1) but not C^2, consistent with the Matérn-$3/2$ sample path regularity (once-differentiable paths almost surely).

Step 2: Posterior mean and variance

With $K_{ij} = k(s_i, s_j) + \tau^2 \delta_{ij}$ the noisy kernel matrix, the GP posterior at a new location s^* is:

$$\mu(s^*) = \mathbf{k}(s^*)^\top (K + \tau^2 I)^{-1} \mathbf{y},$$
$$v(s^*) = k(s^*, s^*) - \mathbf{k}(s^*)^\top (K + \tau^2 I)^{-1} \mathbf{k}(s^*),$$

where $\mathbf{k}(s^*) = (k(s^*, s_1), \dots, k(s^*, s_n))^\top$. The posterior mean $\mu(s^*)$ is the **kriging predictor** (Chapter 11), and $v(s^*)$ is the **kriging variance**.

Step 3: Parameter estimation by marginal likelihood

The **log marginal likelihood** for parameters $\boldsymbol{\theta} = (\sigma^2, \ell, \tau^2)$ is

$$\log p(\mathbf{y} \mid \boldsymbol{\theta}) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \log \det K_{\boldsymbol{\theta}} - \frac{1}{2} \mathbf{y}^\top K_{\boldsymbol{\theta}}^{-1} \mathbf{y},$$

where $K_{\boldsymbol{\theta}} = K(\boldsymbol{\theta}) + \tau^2 I$. Differentiating with respect to σ^2:

$$\frac{\partial \log p}{\partial \sigma^2} = -\frac{1}{2} \operatorname{tr}\!\left(K_{\boldsymbol{\theta}}^{-1} \frac{\partial K}{\partial \sigma^2} \right) + \frac{1}{2} \mathbf{y}^\top K_{\boldsymbol{\theta}}^{-1} \frac{\partial K}{\partial \sigma^2} K_{\boldsymbol{\theta}}^{-1} \mathbf{y}.$$

Since $\partial K / \partial \sigma^2 = K / \sigma^2$, this simplifies to $\frac{\partial \log p}{\partial \sigma^2} = \frac{1}{2\sigma^2} [\mathbf{y}^\top K_{\boldsymbol{\theta}}^{-1} (K/\sigma^2) K_{\boldsymbol{\theta}}^{-1} \mathbf{y} - \operatorname{tr}(K_{\boldsymbol{\theta}}^{-1} K / \sigma^2)]$. Setting to zero gives the profile MLE for σ^2 as a function of ℓ; the length-scale ℓ is estimated by one-dimensional numerical optimisation. In practice this is done jointly by gradient-based optimisation (L-BFGS) using automatic differentiation through the Cholesky factorisation of $K_{\boldsymbol{\theta}}$.

Step 4: Posterior contraction rate

By the Van der Vaart–Van Zanten theorem (Chapter 12), if the true field Z_0 lies in the Matérn-$3/2$ RKHS and the length-scale is tuned as $\ell_n \asymp n^{-1/(2 \cdot 3/2 + 2)} = n^{-1/5}$, the posterior contracts at the minimax rate $n^{-3/2/(3+2)} = n^{-3/7}$ in the $L^2([0,1]^2)$ norm. This is the optimal rate for estimating H^2 functions on $[0,1]^2$.

16.2 Fitting a Log-Gaussian Cox Process to Spatial Point Data

Model and likelihood

Let $\{x_1, \dots, x_n\} \subset W \subset \mathbb{R}^2$ be an observed point pattern in a bounded window W. We model it as a realisation of a **log-Gaussian Cox process** (LGCP):

$$\Phi \mid Z \sim \mathrm{PPP}(e^{Z(\cdot)}), \qquad Z \sim \mathcal{GP}(\mu_0, k),$$

where μ_0 is a constant log-intensity offset and k is a Matérn covariance. The latent log-intensity field Z is unobserved.

Step 1: The intractable likelihood and Laplace approximation

The marginal likelihood integrating over Z is

$$p(\{x_1, \dots, x_n\}) = \mathbb{E}\left[\exp\left(\sum_{i=1}^{n} Z(x_i) - \int_W e^{Z(x)} dx\right)\right],$$

which involves the Laplace functional of the GP and is intractable. The **Laplace approximation** discretises W onto an $m \times m$ grid, approximates Z by its values $\mathbf{z} = (z_1, \dots, z_{m^2})$ at the grid centres, and approximates the integral $\int_W e^{Z(x)} dx \approx |W|/m^2 \sum_j e^{z_j}$. The resulting unnormalised posterior on $\mathbf{z}$ is

$$\log \tilde{p}(\mathbf{z} \mid \{x_i\}) = \sum_{i=1}^{n} z_{c(i)} - \frac{|W|}{m^2} \sum_{j} e^{z_j} - \frac{1}{2} \mathbf{z}^\top K^{-1} \mathbf{z},$$

where $c(i)$ is the grid cell containing x_i. The negative Hessian at the MAP estimate $\mathbf{z}^*$ is

$$W_{\mathbf{z}^*} = \operatorname{diag}\left(\frac{|W|}{m^2} e^{z_j^*}\right) + K^{-1},$$

giving the Laplace approximation $Z \mid \{x_i\} \approx \mathcal{N}(\mathbf{z}^*, W_{\mathbf{z}^*}^{-1})$.

Step 2: MCMC via elliptical slice sampling

For exact inference, **elliptical slice sampling** (Murray–Adams–MacKay, 2010) provides an efficient MCMC algorithm for posteriors of the form $\mathcal{N}(\mathbf{0}, K) \cdot L(\mathbf{z})$ where L is a non-Gaussian likelihood. Each iteration:

1. Draw a "momentum" $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, K)$;
2. Define the ellipse $\mathbf{z}(\theta) = \mathbf{z} \cos\theta + \boldsymbol{\nu} \sin\theta$;
3. Sample θ uniformly from the ellipse via a shrinkage procedure that preserves the likelihood threshold.

The key advantage is that proposals are always on the prior ellipse, so no step-size tuning is needed; the acceptance test only involves the likelihood $L(\mathbf{z}(\theta))$, not the prior. This is particularly efficient for LGCPs where $L(\mathbf{z}) = \exp(\sum z_{c(i)} - |W|/m^2 \sum e^{z_j})$.

Step 3: Intensity estimation and pair correlation

Once samples $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(B)}$ from the posterior are available, the posterior mean intensity is $\hat{\lambda}(x) = B^{-1} \sum_b e^{z^{(b)}(x)}$ and the posterior variance gives a measure of uncertainty. The pair correlation of the LGCP is (Chapter 7)

$$g(r) = e^{k(r)},$$

where $k(r)$ is the covariance kernel evaluated at lag $r = \|s - t\|$. Comparing the empirical pair correlation $\hat{g}(r)$ (estimated via the edge-corrected estimator) with the fitted model $g(r) = e^{\hat{k}(r)}$ provides a goodness-of-fit diagnostic.

16.3 Ripley's K-Function: From PPP to Clustering and Regularity

The canonical comparison

For a stationary isotropic point process Φ on $\mathbb{R}^2$ with intensity λ, the **L-function** (a variance-stabilised version of Ripley's K) is

$$L(r) = \sqrt{\frac{K(r)}{\pi}}.$$

For a homogeneous PPP, $K_{\text{CSR}}(r) = \pi r^2$ so $L_{\text{CSR}}(r) = r$: the L-function grows linearly. Plotting $L(r) - r$ against r gives a diagnostic: values systematically above 0 indicate clustering, values below 0 indicate regularity (inhibition).

Step 1: Theoretical K-functions for canonical models

Thomas process. As computed in Section 7.4, the pair correlation is $g(r) = 1+(4\pi\kappa\sigma^2)^{-1}e^{-r^2/(4\sigma^2)}$, giving

$$K_{\text{Thomas}}(r) = \pi r^2 + \frac{1}{\kappa}\left(1 - e^{-r^2/(4\sigma^2)}\right).$$

The excess over πr^2 is positive and saturates at $1/\kappa$ (the reciprocal of parent density), with the scale of the transition controlled by σ.

Strauss process. For the Strauss process with parameters (γ, R), the pair correlation satisfies $g(r) = \gamma$ for $r \le R$ and $g(r) = 1$ for $r > R$ (to first order in the density). The K-function is thus

$$K_{\text{Strauss}}(r) \approx \gamma\pi r^2 \quad (r \le R), \qquad K_{\text{Strauss}}(r) \approx \pi r^2 - (\pi - \gamma\pi)R^2 \quad (r > R).$$

The deficit $K(r) - \pi r^2 < 0$ for all r reflects the systematic inhibition at short distances.

Step 2: Minimum contrast estimation

Given an observed point pattern, parameters of a model (e.g., κ, σ for the Thomas process) are estimated by minimising the **contrast criterion**

$$C(\kappa, \sigma) = \int_0^{r_{\max}} \left[\hat{K}(r)^q - K_{\text{model}}(r; \kappa, \sigma)^q\right]^2 dr,$$

where $q \in (0, 1)$ is a power chosen to stabilise the estimator (typically $q = 1/4$). The estimator $(\hat{\kappa}, \hat{\sigma}) = \arg\min C$ is consistent and asymptotically normal as $|W| \to \infty$ with $n/|W| \to \lambda$.

Proposition: Asymptotic Normality of Minimum Contrast Estimators

Under regularity conditions on the model $K(r;\boldsymbol{\theta})$ (identifiability, differentiability, existence of moments), the minimum contrast estimator $\hat{\boldsymbol{\theta}}$ satisfies

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\big(\mathbf{0},\ A^{-1}BA^{-1}\big)\,,$$

where $A = \int_0^{r_{\max}} (\partial_{\boldsymbol{\theta}} K^q)(\partial_{\boldsymbol{\theta}} K^q)^\top dr$ is the Fisher-type information matrix and B is the asymptotic variance of $\int (\hat{K}^q - K^q) dr$, estimated by bootstrap or by a moment formula involving the fourth-order cumulant of Φ.

16.4 The Gaussian Free Field: Discrete Approximation and Circle Average

Discrete GFF and convergence

Let $G_n = [-n, n]^2 \cap \mathbb{Z}^2$ be the discrete square lattice with $|V_n| = (2n+1)^2$ vertices, and let Δ_n be the discrete Laplacian: $(\Delta_n f)(x) = \frac{1}{4}\sum_{y\sim x}(f(y) - f(x))$ where the sum is over the 4 lattice neighbours of x. The **discrete GFF** on G_n with Dirichlet boundary conditions is

$$h^n \sim \mathcal{N}(\mathbf{0}, (-\Delta_n)^{-1}),$$

the centred Gaussian with covariance equal to the Green's function $G_n(x,y) = (-\Delta_n)^{-1}(x,y)$ of $-\Delta_n$ on G_n, with $G_n(x,y) = 0$ for x or y on the boundary.

Theorem: Convergence of Discrete GFF to Continuous GFF

Rescale the lattice by $1/n$ so that $G_n/n \subset [-1,1]^2$. For any test function $\varphi \in C_c^\infty((-1,1)^2)$, define the smeared discrete field $h^n(\varphi) = n^{-1}\sum_{x\in G_n} h^n(x)\varphi(x/n)$. Then as $n \to \infty$,

$$h^n(\varphi) \xrightarrow{d} \Phi(\varphi) = \int_{(-1,1)^2} \Phi(x)\varphi(x)\,dx,$$

where Φ is the continuous GFF on $(-1,1)^2$ with Dirichlet boundary conditions. The convergence holds in distribution for all finite-dimensional vectors $(h^n(\varphi_1), \dots, h^n(\varphi_k))$.

The circle average process

Since the GFF Φ is a distribution and not a function, one cannot evaluate it at a single point. The natural regularisation is the **circle average**: for $\varepsilon > 0$ and $z \in D$ with $\operatorname{dist}(z, \partial D) > \varepsilon$, define

$$\Phi_\varepsilon(z) = \frac{1}{2\pi}\int_0^{2\pi} \Phi(z + \varepsilon e^{i\theta})\,d\theta,$$

the average of Φ over the circle of radius ε centred at z.

Proposition: Covariance of the Circle Average

For the GFF on the unit disc $\mathbb{D}$, the circle average satisfies

$$\mathrm{Cov}(\Phi_\varepsilon(z), \Phi_\delta(w)) = \log \frac{1}{\max(\varepsilon, \delta, |z-w|)} + O(1),$$

as $\varepsilon, \delta, |z-w| \to 0$. In particular, $\mathrm{Var}(\Phi_\varepsilon(z)) = \log(1/\varepsilon) + O(1)$, diverging logarithmically as $\varepsilon \to 0$.

Thick points and extrema of the GFF

The points where the GFF is *unusually large* play a central role in the theory of LQG.

Definition: α-Thick Points

For $\alpha > 0$, the **α-thick points** of the GFF are

$$\mathcal{T}_\alpha = \left\{ z \in D : \lim_{\varepsilon \to 0} \frac{\Phi_\varepsilon(z)}{\log(1/\varepsilon)} = \alpha \right\}.$$

Theorem: Hausdorff Dimension of Thick Points (Hu–Miller–Peres, 2010)

Almost surely, $\mathcal{T}_\alpha$ is non-empty for $\alpha \in [0, \sqrt{2/\pi}]$ and empty for $\alpha > \sqrt{2/\pi}$. Its Hausdorff dimension is

$$\dim_H(\mathcal{T}_\alpha) = 2 - \pi\alpha^2/2.$$

The maximum $\alpha^* = \sqrt{2/\pi}$ corresponds to the absolute maximum of the GFF, and $\dim_H(\mathcal{T}_{\alpha^*}) = 0$: the maximisers form a countable (in fact finite) set.

The connection to the LQG measure $M_\gamma(dx) = \varepsilon^{\gamma^2/2} e^{\gamma \Phi_\varepsilon(x)} dx$ is direct: M_γ is supported on the γ-thick points $\mathcal{T}_{\gamma/\pi}$ scaled appropriately, and its Hausdorff dimension $2 - \gamma^2/2$ matches the formula with $\alpha = \gamma/\sqrt{2\pi}$.

16.5 Determinantal Point Processes for Diverse Subset Selection

The subset selection problem

Given a set $\mathcal{Y} = \{1, \dots, N\}$ of items, each represented by a feature vector $\phi_i \in \mathbb{R}^d$, we wish to select a **diverse** subset $Y \subseteq \mathcal{Y}$ that covers the feature space without redundancy. Determinantal point processes provide a principled probabilistic model for this: subsets with high diversity (high determinant of the feature Gram matrix) are assigned higher probability.

Step 1: The quality-diversity decomposition

Define the **quality-diversity kernel** $L_{ij} = q_i \, \phi_i^\top \phi_j \, q_j$ where $q_i > 0$ is a **quality score** and $\phi_i \in \mathbb{R}^d$ with $\|\phi_i\| = 1$ is a normalised feature vector. The L-ensemble DPP assigns

$$\mathbb{P}(Y) = \frac{\det L_Y}{\det(I + L)}.$$

The determinant $\det L_Y = \prod_{i \in Y} q_i^2 \cdot \det[\phi_i^\top \phi_j]_{i,j \in Y}$ factorises into a quality term (product of quality scores) and a diversity term (volume squared of the parallelepiped spanned by $\{\phi_i\}_{i \in Y}$). The L-ensemble therefore automatically trades off quality against diversity.

Step 2: The normalisation constant

Proposition: Partition Function of the L-Ensemble

$\det(I + L) = \sum_{Y \subseteq \mathcal{Y}} \det L_Y$, where the sum is over all 2^N subsets of $\mathcal{Y}$. If $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_N$ are the eigenvalues of L, then

$$\det(I + L) = \prod_{n=1}^{N} (1 + \mu_n).$$

Proof. The matrix determinant lemma gives $\det(I + L) = \det\big[\sum_n (1 + \mu_n) v_n v_n^\top / (1 + \mu_n) + \mu_n / (1 + \mu_n)\big]$. More directly, by expanding the determinant: $\det(I + L) = \sum_{S \subseteq [N]} \det(L_S)$ is the sum of all principal minors of L, which equals $\prod_n (1 + \mu_n)$ by the spectral theorem. □

Step 3: Sampling algorithm and complexity

By the eigendecomposition $L = V \operatorname{diag}(\mu_1, \ldots, \mu_N) V^\top$, the exact sampling algorithm from Section 10.4 runs in two phases:

Phase 1. For each eigenvalue μ_n, include eigenvector v_n in a set J with probability $\mu_n / (1 + \mu_n)$. The expected size of the selected subset is $\mathbb{E}[|Y|] = \sum_n \mu_n / (1 + \mu_n)$.

Phase 2. Given $|J| = k$, greedily select k items from $\{1, \ldots, N\}$ as follows: at each step, add item i with probability proportional to $\det[L_{Y \cup \{i\}}] / \det[L_Y]$, which equals the squared distance from ϕ_i to the subspace spanned by $\{\phi_j\}_{j \in Y}$ (by the matrix determinant lemma). This is equivalent to a Gram–Schmidt orthogonalisation in feature space.

Example: Document Summarisation via DPP

In extractive document summarisation, $\mathcal{Y}$ is the set of sentences, ϕ_i is the normalised TF-IDF feature vector of sentence i, and q_i is a quality score (e.g., relevance to a query). The DPP selects a summary of sentences that are individually relevant (q_i large) and mutually dissimilar ($\phi_i^\top \phi_j$ small for selected pairs $i \neq j$).

Empirically, DPP-based summaries outperform greedy relevance-only methods by achieving better diversity with comparable relevance. The partition function $\det(I + L)$ serves as a principled measure of the "total diversity budget" of the document collection.

Remark: Continuous DPPs for Spatial Coverage

For a spatial coverage problem (deploying sensors in $[0,1]^2$ to maximally cover the domain), the DPP on $\mathcal{X} = [0,1]^2$ with the squared-exponential kernel $K(x,y) = e^{-\|x-y\|^2/(2\ell^2)}$ models a repulsive placement that avoids redundant nearby sensors. The expected number of sensors placed is $\int K(x,x)dx = 1$ (since $K(x,x) = 1$), so to place n sensors one scales the kernel to nK. The resulting process is repulsive at scale ℓ and approximately uniform at scales $\gg \ell$, combining the Poisson randomness of large-scale placement with DPP repulsion at small scales.

Chapter 17

Permanental Point Processes and Bosonic Systems

17.1 Definition and Basic Properties

Determinantal point processes model the repulsive statistics of fermions; their counterpart, *permanental point processes*, model the attractive (or at least non-repulsive) statistics of bosons. Where a DPP has correlation functions given by determinants of a kernel matrix, a permanental process has correlation functions given by *permanents*, encoding the positive correlations characteristic of bosonic bunching.

Definition: Permanental Point Process

A point process Φ on $(\mathcal{X}, \mu)$ is a **permanental point process** (PmPP) with kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ and index $\alpha > 0$ if, for all $n \geq 1$ and pairwise distinct $x_1, \dots, x_n \in \mathcal{X}$,

$$\rho_n(x_1, \dots, x_n) = \alpha^n \cdot \operatorname{perm}[K(x_i, x_j)/\alpha]_{1 \leq i,j \leq n},$$

where the **permanent** of an $n \times n$ matrix A is

$$\operatorname{perm}(A) = \sum_{\pi \in S_n} \prod_{i=1}^{n} A_{i,\pi(i)}.$$

The Poisson process is recovered in the limit where $K(x, y) \to 0$ for $x \neq y$ and $K(x, x) \to \lambda(x)$; in this case all permanents factor and one recovers the Poisson correlation functions.

The permanent satisfies $\operatorname{perm}(A) \geq 0$ for non-negative matrices A, so permanental processes are well-defined as positive measures. This contrasts sharply with DPPs, where the antisymmetric determinant can nearly vanish when points approach each other. The $\alpha = 1$ case, due to Macchi (1971) in the bosonic context, corresponds to the "boson process" of quantum statistical mechanics.

Theorem: Existence of Permanental Point Processes (Vere-Jones)

A permanental point process with kernel K and index $\alpha > 0$ exists on $\mathcal{X}$ if and only if the integral operator with kernel K on $L^2(\mathcal{X}, \mu)$ has all eigenvalues $\lambda_n \geq 0$ and

$$\sum_n \lambda_n < \infty \quad \text{(trace-class condition)}.$$

When this holds, the void probability is $\mathbb{P}(\Phi(B) = 0) = \det(I - K_B)^{-\alpha}$ for bounded Borel sets B, where K_B is the restriction of K to $L^2(B, \mu)$.

Sketch. The key observation is that the Laplace functional of a permanental process has the form

$$\mathcal{L}_\Phi(f) = \mathbb{E}\big[e^{-\Phi(f)}\big] = \det(I + \tilde{K}_f)^{-\alpha},$$

where $\tilde{K}_f$ is the operator with kernel $K(x,y)(e^{f(x)} - 1)^{1/2}(e^{f(y)} - 1)^{1/2}$. The expansion of $\det(I + \tilde{K}_f)^{-\alpha}$ as a power series in the entries of $\tilde{K}_f$ (via the identity $\log\det = \operatorname{tr}\log$) recovers the permanental correlation functions, provided the series converges, which requires the trace-class condition. □

Example: The Boson Process on $\mathbb{R}^d$

Take $\mathcal{X} = \mathbb{R}^d$, $\mu =$ Lebesgue measure, $\alpha = 1$, and the kernel

$$K(x,y) = \lambda\, e^{-\|x-y\|^2/(2\sigma^2)},$$

a scaled Gaussian kernel. The intensity is $K(x,x) = \lambda$ (homogeneous), and the pair correlation function is

$$g(r) = 1 + e^{-r^2/\sigma^2} \geq 1.$$

This exhibits *clustering*: pairs of points at distances $r \lesssim \sigma$ are more common than in a Poisson process, while for $r \gg \sigma$ the process is approximately Poisson. The parameter σ controls the bunching scale, analogous to the coherence length in a Bose–Einstein condensate. The trace-class condition $\int K(x,x)\,dx = \lambda \cdot \mathrm{Vol}(\mathcal{X})$ requires a finite domain or a decaying kernel.

17.2 Cox Representation and Mixture Structure

Every permanental point process admits a representation as a mixture of Poisson processes, providing an intuitive counterpart to the determinantal structure of DPPs.

Theorem: Cox Representation of Permanental Processes

Let Φ be a permanental process with kernel K and index $\alpha > 0$. Then Φ has the same distribution as a Cox process directed by the random intensity measure

$$\Lambda(dx) = \sum_{j=1}^{\infty} \Gamma_j \cdot |\phi_j(x)|^2 \, d\mu(x),$$

where (λ_j, ϕ_j) are the eigenvalue–eigenfunction pairs of the operator T_K, and the $\Gamma_j \sim \text{Gamma}(\alpha, 1/\lambda_j)$ are independent Gamma random variables.

This representation reveals the structure of permanental clustering: the driving intensity Λ is a random mixture of the squared eigenfunctions $|\phi_j|^2$, each modulated by a Gamma weight. Since the Gamma distribution has a heavier tail than Poisson, the resulting process has larger fluctuations in local density than a Poisson process, consistent with bosonic bunching.

17.3 Applications in Machine Learning: Repulsion versus Attraction

Both DPPs and permanental processes arise in the design of sampling schemes for machine learning. DPP samples provide diverse, repulsive subsets useful for data summarisation, while permanental samples model clustered selection strategies useful for multi-task learning and ensemble methods.

Remark: Generative Models via Point Processes

In a deep generative model context, a latent DPP over feature space ensures that sampled latent codes are diverse (spread across the representation space), regularising against mode collapse in variational autoencoders. Conversely, a permanental prior over the index set of a neural network's attention heads encourages heads to cluster on similar query regions, providing a probabilistic framework for group attention mechanisms. The permanent and determinant thus serve as complementary regularisers encoding two opposing geometric priors on sets.

Chapter 18

Random Matrices and Spatial Eigenvalue Distributions

18.1 The Wigner Semicircle Law

Random matrix theory studies the statistical properties of eigenvalues of large matrices with random entries. It connects to spatial randomness through the DPP structure of eigenvalue processes and the universality of limiting spectral measures.

Definition: Wigner Matrix Ensemble

A **Wigner matrix** of size n is a random symmetric (or Hermitian) matrix W_n whose upper-triangular entries $\{W_{ij} : i \le j\}$ are independent real-valued random variables with $\mathbb{E}[W_{ij}] = 0$, $\mathbb{E}[W_{ij}^2] = \sigma^2/n$ for $i < j$, and $\mathbb{E}[W_{ii}^2] = \sigma_d^2/n$. The normalisation by $1/n$ ensures that the empirical spectral measure $\mu_n = n^{-1}\sum_{k=1}^{n}\delta_{\lambda_k}$ has a non-trivial weak limit.

Theorem: Wigner Semicircle Law

Let W_n be a Wigner matrix with variance σ^2 and assume finite fourth moments. As $n \to \infty$, the empirical spectral measure $\mu_n = n^{-1}\sum_{k}\delta_{\lambda_k(W_n)}$ converges weakly in probability to the **semicircle distribution**

$$\mu_{\mathrm{sc}}(dx) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - x^2}\,\mathbf{1}_{|x|\le 2\sigma}\,dx.$$

Proof via the method of moments. By the **moment method**, it suffices to show that $\mathbb{E}[\mathrm{tr}(W_n^k)/n] \to \int x^k \mu_{\mathrm{sc}}(dx)$ for all $k \ge 1$. Expanding $\mathrm{tr}(W_n^k) = \sum_{i_1,\dots,i_k} W_{i_1 i_2} W_{i_2 i_3}\cdots W_{i_k i_1}$, the expectation is non-zero only when every entry $W_{i_j i_{j+1}}$ appears an even number of times (by independence and zero mean). Such index sequences $i_1,\dots,i_k,i_1$ are in bijection with non-crossing pair partitions of $\{1,\dots,k\}$ (by Wigner's combinatorial analysis, 1955), and the number of such partitions is the Catalan number $C_{k/2}$ for even k and 0 for odd k. Since $\int x^k\mu_{\mathrm{sc}}(dx) = C_{k/2}\sigma^k$ (a standard integral identity for the semicircle), the two sequences agree, completing the proof. $\square$

Remark: Universality of the Semicircle Law

The Wigner semicircle law holds under minimal moment assumptions (finite second moment suffices for the weak convergence, and finite fourth moment for the convergence in probability): the limiting spectral distribution depends only on the variance σ^2, not on the precise distribution of the entries. This **universality** is an analogue of the central limit theorem for matrix spectra. At the scale of individual eigenvalue spacings (the "local" level), the GUE sine kernel also appears universally, a deeper result requiring much more delicate analysis (Erdős–Schlein–Yau, 2009–2012).

18.2 The Marchenko–Pastur Law for Sample Covariance Matrices

In statistical applications, one often encounters sample covariance matrices rather than Wigner matrices. Their spectral distribution obeys the Marchenko–Pastur law, which plays a fundamental role in high-dimensional statistics.

Definition: Sample Covariance Matrix

Let X be an $n \times p$ matrix with i.i.d. entries of mean zero and variance $1/p$. The **sample covariance matrix** is $S_n = X^\top X \in \mathbb{R}^{p \times p}$, a positive semi-definite matrix whose eigenvalues are all non-negative.

Theorem: Marchenko–Pastur Law

Suppose $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Then the empirical spectral measure $\mu_n = p^{-1} \sum_{k=1}^{p} \delta_{\lambda_k(S_n)}$ converges weakly in probability to the **Marchenko–Pastur distribution** $\mu_{\mathrm{MP},\gamma}$ with density

$$\rho_\gamma(x) = \frac{1}{2\pi x} \sqrt{(\lambda_+ - x)(x - \lambda_-)}\, \mathbf{1}_{[\lambda_-, \lambda_+]}(x),$$

where $\lambda_\pm = (1 \pm \sqrt{\gamma})^2$. If $\gamma > 1$, there is additionally a point mass $(1 - 1/\gamma)$ at 0.

The Marchenko–Pastur law has immediate implications for the eigenvalue structure of sample covariance matrices in high dimensions. When $\gamma = p/n$ is small (many more observations than variables), the eigenvalues cluster tightly near 1 and the sample covariance is a reliable estimator. When γ approaches 1 (as in many modern genomic or financial datasets), the eigenvalues spread dramatically over the interval $[\lambda_-, \lambda_+]$, with the top eigenvalue near $\lambda_+ = (1+\sqrt{\gamma})^2 \gg 1$. This **eigenvalue spreading** is a form of noise amplification inherent in high-dimensional estimation, independent of the signal structure.

Proposition: Estimation of the True Spectrum via the Marchenko–Pastur Law

Suppose the true covariance is $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2)$ and $X = Z\Sigma^{1/2}$ where Z has i.i.d. entries. The empirical spectral measure of $S_n = X^\top X/p$ converges to the free convolution $\mu_{\mathrm{MP},\gamma} \boxtimes \nu$ where $\nu = p^{-1}\sum_k \delta_{\sigma_k^2}$ is the empirical spectral measure of Σ. Inverting this relation via the **Marchenko–Pastur equation** for the Stieltjes transform

$$m(z) = \int \frac{d\nu(\sigma^2)}{\sigma^2(1-\gamma-\gamma z m(z)) - z}$$

allows estimation of ν (the true spectrum) from the observed μ_n via a fixed-point iteration, forming the basis of optimal shrinkage estimators (Ledoit–Wolf, 2004; El Karoui, 2008).

18.3 Beta Ensembles and Tridiagonal Models

The GUE ($\beta = 2$), GOE ($\beta = 1$), and GSE ($\beta = 4$) are special cases of a one-parameter family of random matrix ensembles indexed by $\beta > 0$, with joint eigenvalue densities proportional to $\prod_{i<j} |\lambda_i - \lambda_j|^\beta \prod_i e^{-\lambda_i^2/2}$.

Theorem: Tridiagonal Representation of Beta Ensembles (Dumitriu–Edelman)

The eigenvalue distribution of the β-ensemble (for any $\beta > 0$) is that of the tridiagonal symmetric matrix

$$H_n = \frac{1}{\sqrt{n\beta}} \begin{pmatrix} N_1 & \chi_{(n-1)\beta} & & \\ \chi_{(n-1)\beta} & N_2 & \chi_{(n-2)\beta} & \\ & \ddots & \ddots & \ddots \\ & & \chi_\beta & N_n \end{pmatrix},$$

where $N_i \sim \mathcal{N}(0,2)$ i.i.d. and $\chi_{k\beta}$ is a chi-distributed random variable with $k\beta$ degrees of freedom, all independent. The Wigner semicircle law holds for all $\beta > 0$.

Chapter 19

Spatial Extremes and Max-Stable Processes

19.1 Extreme Value Theory for Spatial Fields

The study of the maximum of a random field over a domain is of fundamental importance in environmental science (extreme precipitation, temperatures, wind speeds), materials science (failure at the weakest site), and finance (portfolio tail risk). Max-stable processes are the natural limit objects for the properly normalised maxima of i.i.d. copies of a random field.

Definition: Max-Stable Process

A stochastic process $\{Z(x) : x \in T\}$ with $Z(x) > 0$ a.s. is **max-stable** if, for every $n \geq 1$, the finite-dimensional distributions of $n^{-1}\max_{i=1}^{n} Z^{(i)}(x)$ (where $Z^{(1)}, \dots, Z^{(n)}$ are i.i.d. copies) agree with those of $Z(x)$:

$$\mathbb{P}\left(\frac{1}{n}\max_{i=1}^{n} Z^{(i)}(x_j) \leq z_j,\ j = 1, \dots, k\right) = \mathbb{P}(Z(x_j) \leq z_j,\ j = 1, \dots, k).$$

Theorem: Spectral Representation of Max-Stable Processes (de Haan)

Every max-stable process with unit-Fréchet marginals $\mathbb{P}(Z(x) \leq z) = e^{-1/z}$ admits a **spectral representation**

$$Z(x) = \max_{i \geq 1} U_i\, f_i(x),$$

where $\{U_i\}$ are the points of a unit-rate Poisson process on $(0, \infty)$ (so $U_i \sim \mathrm{PPP}(\{u^{-2}du\})$) and $\{f_i\}$ are i.i.d. non-negative random functions (the **spectral functions**) with $\mathbb{E}[f(x)] = 1$ for all x. The finite-dimensional exponent measure is determined by $\mathbb{E}[\max_i U_i f_i(x)]$.

Derivation. The key identity for unit-Fréchet marginals is $\mathbb{P}(Z(x) \leq z) = \exp(-\mathbb{E}[f(x)]/z) =$

$e^{-1/z}$, which requires $\mathbb{E}[f(x)] = 1$. The joint survival probability is:

$$\mathbb{P}(Z(x_j) \le z_j, j = 1, \dots, k) = \exp\left(-\mathbb{E}\left[\max_{j=1}^{k} \frac{f(x_j)}{z_j}\right]\right) =: e^{-V(z_1,\dots,z_k)},$$

where V is the **exponent function**, a homogeneous function of degree -1. The spectral representation makes V explicit: $V(z_1, \dots, z_k) = \mathbb{E}[\max_j f(x_j)/z_j]$. □

19.2 The Smith Model and the Brown–Resnick Process

Definition: Smith Model

The **Smith model** on $\mathbb{R}^d$ is the max-stable process obtained by taking spectral functions $f_i(x) = \phi(x - c_i; \Sigma)$ where $\phi(\cdot; \Sigma)$ is a multivariate Gaussian density with covariance Σ and $\{c_i\}$ are the points of a unit-rate PPP on $\mathbb{R}^d$. The resulting process has exponent function

$$V(z_1, \dots, z_k) = \sum_{j=1}^{k} \frac{1}{z_j} \Phi_d\left(\mathbf{a}_j + \frac{1}{2}\mathbf{v}_j;\, \Sigma_j\right),$$

where Φ_d is the d-variate Gaussian CDF and the vectors $\mathbf{a}_j, \mathbf{v}_j$ depend on the geometry of the configuration $(x_1, \dots, x_k)$ and the covariance Σ.

Definition: Brown–Resnick Process

The **Brown–Resnick process** is the max-stable process with spectral functions

$$f_i(x) = e^{W_i(x) - \sigma^2(x)/2},$$

where W_i are i.i.d. copies of a zero-mean Gaussian process with variance $\sigma^2(x) = \mathrm{Var}(W(x))$ and stationary increments (a generalised Brownian motion on $\mathbb{R}^d$). When W is fractional Brownian motion with Hurst index H, the variogram $2\gamma(h) = \mathbb{E}[(W(x+h) - W(x))^2] = \|h\|^{2H}$ determines the spatial dependence structure.

Theorem: Brown–Resnick as a Limit of Rescaled Gaussian Fields

Let $\{X^{(i)}(x)\}_{i=1}^{n}$ be i.i.d. copies of a stationary Gaussian field with mean $-\sigma^2(x)/2$ and variogram $\gamma(h)$. The pointwise maximum, properly normalised, satisfies

$$\frac{\max_{i=1}^{n} X^{(i)}(x)}{\sigma(0)} - b_n \xrightarrow{d} Z(x), \quad n \to \infty,$$

where $b_n = \sqrt{2 \log n}$ and Z is the Brown–Resnick process associated to γ. This establishes the Brown–Resnick process as the universal extreme-value limit for Gaussian random fields under appropriate domain-growing conditions.

19.3 Inference for Max-Stable Processes

Statistical inference for max-stable processes presents challenges absent from the Gaussian case: the likelihood is not available in closed form for more than a handful of sites. Composite likelihood methods, which replace the full likelihood by a product of lower-dimensional marginal likelihoods, provide a tractable alternative.

Definition: Pairwise Composite Likelihood

For observations $\mathbf{z}_t = (Z_t(x_1), \dots, Z_t(x_p))$, $t = 1, \dots, n$ of a max-stable process at p sites, the **pairwise composite log-likelihood** is

$$\ell_C(\boldsymbol{\theta}) = \sum_{t=1}^{n} \sum_{1 \le j < k \le p} \log f_{jk}(Z_t(x_j), Z_t(x_k); \boldsymbol{\theta}),$$

where $f_{jk}(z_j, z_k; \boldsymbol{\theta})$ is the bivariate density of the process at (x_j, x_k) under parameter $\boldsymbol{\theta}$. The MCLE $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_C(\boldsymbol{\theta})$ is consistent and asymptotically normal with sandwich covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\theta}}) \approx H^{-1} J H^{-1}$ where $H = -\mathbb{E}[\nabla^2 \ell_C]$ and $J = \mathrm{Var}(\nabla \ell_C)$.

Proposition: Bivariate Density of the Brown–Resnick Process

For the Brown–Resnick process at two sites x, y with variogram value $2\gamma = 2\gamma(\|x - y\|)$, the bivariate density of $(Z(x), Z(y)) = (z_1, z_2)$ under unit-Fréchet marginals is:

$$f(z_1, z_2) = \exp\!\left(-\frac{\Phi(a_1)}{z_1} - \frac{\Phi(a_2)}{z_2}\right) \times \left[\left(\frac{\phi(a_1)}{z_1^2 \sqrt{2\gamma}} + \frac{\Phi(a_1)}{z_1^2}\right)\left(\frac{\phi(a_2)}{z_2^2 \sqrt{2\gamma}} + \frac{\Phi(a_2)}{z_2^2}\right) + \frac{\phi(a_1)\phi(a_2)}{4\gamma z_1 z_2}\right],$$

where $a_1 = \sqrt{2\gamma}/2 + \log(z_1/z_2)/\sqrt{2\gamma}$, $a_2 = \sqrt{2\gamma}/2 + \log(z_2/z_1)/\sqrt{2\gamma}$, and ϕ, Φ are the standard Gaussian density and CDF.

Chapter 20

Gaussian Processes on Graphs and Combinatorial Spaces

20.1 Heat Kernels on Graphs

Random fields indexed by vertices of a graph arise naturally in network data analysis, genomics (gene regulatory networks), social science (social network data), and signal processing (graph neural networks). The heat kernel of the graph Laplacian plays the role of the squared-exponential kernel in the Euclidean setting.

Definition: Graph Laplacian and Heat Kernel

Let $G = (V, E, w)$ be a weighted undirected graph with vertex set $V = \{1, \dots, p\}$, edges E, and non-negative weights w_{ij}. The **weighted Laplacian** is the $p \times p$ matrix $L = D - W$, where $W_{ij} = w_{ij}$ and $D = \operatorname{diag}(\sum_j w_{ij})$ is the degree matrix. The **heat kernel** at diffusion time $t > 0$ is

$$H_t = e^{-tL} = \sum_{k=0}^{\infty} \frac{(-t)^k L^k}{k!} = \sum_{\ell=0}^{p-1} e^{-t\lambda_\ell}\, u_\ell u_\ell^{\top},$$

where $0 = \lambda_0 \le \lambda_1 \le \cdots \le \lambda_{p-1}$ and $\{u_\ell\}$ are the eigenvalues and eigenvectors of L.

The heat kernel H_t is positive semi-definite for every $t \ge 0$, as follows from the spectral expansion: $e^{-t\lambda_\ell} \ge 0$. It therefore defines a valid covariance kernel for a Gaussian random field on V.

Proposition: Properties of the Graph Heat Kernel

The heat kernel $H_t(i, j) = (e^{-tL})_{ij}$ satisfies:

(i) **Semigroup property**: $H_{s+t} = H_s H_t$;

(ii) **Boundary condition**: $H_0 = I$ (identity matrix);

(iii) **Diffusion**: $\frac{d}{dt} H_t = -L H_t$;

(iv) **Stochastic**: $\sum_j H_t(i,j) = e^{-t\lambda_0} = 1$ (row sums equal 1 since $\lambda_0 = 0$ and $u_0 = \mathbf{1}/\sqrt{p}$);

(v) **Decay with distance**: $H_t(i,j)$ decreases as the graph distance $d_G(i,j)$ increases, with the rate governed by t (larger t = more diffusion = longer-range correlations).

Example: Heat Kernel on the d-Dimensional Torus

On the discrete d-dimensional torus $\mathbb{Z}_n^d$ (the Cayley graph of $(\mathbb{Z}/n\mathbb{Z})^d$ with generators $\pm e_1, \ldots, \pm e_d$), the eigenvectors of L are Fourier modes $u_k(x) = e^{2\pi i k \cdot x/n}/n^{d/2}$ with eigenvalues $\lambda_k = 2d - 2\sum_{j=1}^d \cos(2\pi k_j/n)$. The heat kernel is thus

$$H_t(x,y) = \frac{1}{n^d} \sum_{k \in \mathbb{Z}_n^d} \exp\left(-t\left(2d - 2\sum_j \cos\frac{2\pi k_j}{n}\right)\right) e^{2\pi i k\cdot(x-y)/n}.$$

As $n \to \infty$ and $t \to \infty$ with $t/n^2 \to s$, this converges to the heat kernel of the flat torus $[0,1]^d$: $H_s(x,y) = \sum_{k\in\mathbb{Z}^d} e^{-4\pi^2 s\|k\|^2} e^{2\pi i k\cdot(x-y)}$, demonstrating the convergence of the graph GP to the continuum GP on the torus.

20.2 Matérn-Type Kernels via Graph Differential Operators

The Matérn family of kernels on $\mathbb{R}^d$, defined spectrally by $\hat{k}_\nu(\omega) = (\kappa^2 + \|\omega\|^2)^{-(\nu+d/2)}$, can be extended to graphs by replacing the Fourier operator $\|\omega\|^2$ with the graph Laplacian L.

Definition: Graph Matérn Kernel

For smoothness parameter $\nu > 0$ and scale $\kappa > 0$, the **graph Matérn kernel** is the $p \times p$ positive semi-definite matrix

$$K_\nu = \sigma^2 \left(\frac{2\nu}{\kappa^2} I + L\right)^{-\nu - p/2},$$

where I is the $p \times p$ identity matrix and σ^2 is a variance parameter. In the spectral domain, $(K_\nu)_{\ell\ell} = \sigma^2(2\nu/\kappa^2 + \lambda_\ell)^{-\nu-p/2}$, directly generalising the Matérn spectral density.

For integer $\nu = 1/2, 3/2, 5/2$, the graph Matérn kernel can be expressed as a (finite) polynomial of the resolvent $(2\nu/\kappa^2 I + L)^{-1}$ times a lower-order power, making computation via sparse Cholesky factorisation of the graph Laplacian (which is sparse) feasible even for $p \sim 10^6$ nodes.

Theorem: SPDE Characterisation of Graph Matérn Fields

A Gaussian random field $f : V \to \mathbb{R}$ on a graph with precision matrix $\Omega = \sigma^{-2}(2\nu/\kappa^2 I + L)^{\nu+p/2}$ is the unique solution to the graph pseudo-differential equation

$$\left(\frac{2\nu}{\kappa^2} I + L\right)^{(\nu+p/2)/2} f = \sigma^{-1}\mathcal{W},$$

where $\mathcal{W} \sim \mathcal{N}(\mathbf{0}, I)$ is graph white noise. This provides the graph analogue of the Whittle–Lindgren–Rue SPDE characterisation of Matérn fields on $\mathbb{R}^d$.

20.3 Signal Processing on Graphs: Graph Fourier Transform

The graph Fourier transform (GFT) provides a spectral decomposition of functions on V analogous to the classical Fourier transform, with the eigenvectors of L playing the role of complex exponentials.

Definition: Graph Fourier Transform

Let $L = U\Lambda U^\top$ be the eigendecomposition of the graph Laplacian, with $U = [u_0 | \cdots | u_{p-1}]$ the matrix of eigenvectors and $\Lambda = \operatorname{diag}(\lambda_0, \ldots, \lambda_{p-1})$ the diagonal matrix of eigenvalues. The **graph Fourier transform** of a signal $f : V \to \mathbb{R}$ (viewed as a vector $\mathbf{f} \in \mathbb{R}^p$) is

$$\hat{\mathbf{f}} = U^\top \mathbf{f} \in \mathbb{R}^p, \qquad \hat{f}(\ell) = \sum_{i \in V} f(i) u_\ell(i).$$

The inverse GFT is $\mathbf{f} = U\hat{\mathbf{f}} = \sum_\ell \hat{f}(\ell) u_\ell$. The Parseval identity holds: $\|\mathbf{f}\|^2 = \|\hat{\mathbf{f}}\|^2$.

Proposition: Spectral Convolution on Graphs

For a graph signal $\mathbf{f}$ and a filter $\mathbf{h}$ (another signal or a function of the eigenvalues), the **spectral graph convolution** is defined as

$$\mathbf{f} *_G \mathbf{h} = U(\hat{\mathbf{f}} \odot \hat{\mathbf{h}}) = U \operatorname{diag}(h(\lambda_0), \ldots, h(\lambda_{p-1}))\, U^\top \mathbf{f},$$

where $\odot$ denotes elementwise (Hadamard) product and $h : \mathbb{R} \to \mathbb{R}$ is a filter function applied to the eigenvalues. This operation is the graph analogue of the classical convolution theorem $\widehat{f * g} = \hat{f} \cdot \hat{g}$.

Example: Graph Convolutional Networks as GP Approximations

A single-layer **graph convolutional network** (GCN) with spectral filter $h(\lambda) = 1 - \lambda/(2\lambda_{\max})$ (the Chebyshev approximation of order 1 to a low-pass filter) implements the mapping $\mathbf{f} \mapsto (I - L/(2\lambda_{\max}))\mathbf{f}$. This is equivalent to a graph GP

regression step with a truncated heat kernel prior, smoothing the input signal across neighbouring nodes. Multi-layer GCNs correspond to multiple applications of this low-pass filter, progressively aggregating information over larger graph neighbourhoods. The connection to GP regression allows Bayesian uncertainty quantification for GCN predictions via the posterior covariance of the corresponding GP.

Chapter 21

Gaussian Free Field: Further Topics

21.1 The Discrete GFF and the Random Walk Connection

The discrete GFF on a finite graph (V, E) with boundary ∂V is intimately connected to simple random walks. This connection provides probabilistic intuition for the Green's function covariance.

Proposition: GFF Covariance and Hitting Probabilities

Let $G_n(x, y) = (-\Delta_n)^{-1}(x, y)$ be the Green's function of the discrete Laplacian on $G_n = [-n, n]^2 \cap \mathbb{Z}^2$ with Dirichlet boundary conditions. Then

$$G_n(x, y) = \mathbb{E}_x[\text{number of visits to } y \text{ before hitting } \partial G_n]\,,$$

the expected number of visits to y by a simple random walk started at x, before the walk first reaches the boundary ∂G_n. Consequently, the covariance $\mathrm{Cov}(h^n(x), h^n(y)) = G_n(x, y)$ of the discrete GFF equals the expected number of times a walk from x visits y before exiting.

Proof. By definition of the Green's function as the inverse of the negative discrete Laplacian, $G_n(x, y) = (-\Delta_n)^{-1}(x, y)$. The resolvent formula for the discrete Laplacian gives $(-\Delta_n)^{-1}(x, y) = \sum_{t=0}^{\infty} P^t(x, y)$, where $P^t(x, y)$ is the t-step transition probability of the simple random walk on G_n, killed at the boundary. Summing over t yields the expected number of visits to y before absorption, as stated. □

This probabilistic representation of the GFF covariance has several consequences. First, it explains why the GFF variance $G_n(x, x)$ grows logarithmically as x moves away from the boundary: a random walk started at x takes $O(\log n)$ steps on average to reach ∂G_n in $d = 2$, and visits x a number of times proportional to this. Second, the joint law of the GFF can be sampled by simulating random walk local times:

Theorem: Isomorphism Theorem (Dynkin, 1984)

Let $\{L_t^x : x \in V, t \geq 0\}$ be the local times of a continuous-time simple random walk on a finite graph G with boundary ∂G. Let τ_∂ be the hitting time of ∂G. Then, conditionally on the local times $\{L_{\tau_\partial}^x\}_{x\in V}$, the $h^2/2$ **process** has the same law as $\frac{1}{2}\{h(x)^2\}$ where $\{h(x)\}_{x\in V}$ is a discrete GFF on G:

$$\{L_{\tau_\partial}^x : x \in V\} \overset{d}{=} \left\{\frac{h(x)^2}{2} + L_{\tau_\partial,h}^x : x \in V\right\},$$

where the right side involves the GFF squared local times. This isomorphism theorem links the distributional properties of random walk local times to the GFF, and is a key tool in understanding the geometry of the GFF level sets.

21.2 Level Sets of the GFF and the Conformally Invariant Contours

A central theme in modern probability is the identification of conformally invariant structures in two-dimensional critical models. The zero level set of the GFF—the random curve where $\Phi(z) = 0$—turns out to be closely related to the Schramm–Loewner Evolution with parameter $\kappa = 4$.

Definition: Schramm–Loewner Evolution SLE_κ

For $\kappa > 0$, the **SLE_κ** is a random fractal curve $\gamma : [0, \infty) \to \overline{\mathbb{H}}$ in the upper half-plane generated by the Loewner equation

$$\partial_t g_t(z) = \frac{2}{g_t(z) - W_t}, \qquad g_0(z) = z,$$

where the **driving function** $W_t = \sqrt{\kappa}\, B_t$ is $\sqrt{\kappa}$ times a standard Brownian motion. The curve γ is the "trace" of the growing hull $K_t = \overline{\mathbb{H}} \setminus g_t^{-1}(\mathbb{H})$.

Theorem: GFF Level Lines and SLE_4 (Schramm–Sheffield)

The zero-height level lines of the GFF Φ on the upper half-plane $\mathbb{H}$, defined as limits of the zero-contours of smooth approximations Φ_ε, are distributed as SLE_4 curves. More precisely, the level set $\{z : \Phi(z) = 0\}$ is (a.s.) a countable union of SLE_4-type curves in $\mathbb{H}$ with Dirichlet boundary conditions. The Hausdorff dimension of these curves is $\dim_H(\{z : \Phi(z) = 0\}) = 1 + 4/8 = 3/2$ a.s.

This theorem, combined with the Duplantier–Sheffield theorem on Liouville quantum gravity, establishes the GFF as the universal "master field" encoding both the geometry (via SLE level lines) and the measure (via LQG) of two-dimensional random geometry.

21.3 The Gaussian Free Field in Higher Dimensions

While the $d = 2$ GFF has a rich conformal structure, the GFF in $d \geq 3$ has qualitatively different properties: it is a genuine function (in H^1 sense) rather than a distribution, and its level sets are $(d-1)$-dimensional hypersurfaces.

Proposition: Regularity of the GFF in $d \geq 3$

In $d \geq 3$, the GFF Φ on a bounded domain $D \subset \mathbb{R}^d$ with Green's function $G_D(x, y) \sim c_d \|x - y\|^{-(d-2)}$ lies almost surely in the Sobolev space $H^1(D)$ (the space of $L^2(D)$ functions with L^2 gradients). In particular, Φ is a function (not merely a distribution) for $d \geq 3$, and its sample paths satisfy $\mathbb{E}[\|\nabla\Phi\|^2_{L^2(D)}] = \mathrm{tr}(-\Delta_D^{-1} \cdot (-\Delta_D)) = \mathrm{tr}(I) = (\text{dimension of } L^2(D))$, which is finite in an appropriately regularised sense.

The capacity theory of the $d \geq 3$ GFF connects its level-set geometry to the potential theory of Brownian motion: the capacity of a set A relative to G_D governs the probability that the GFF level set $\{\Phi = h\}$ intersects A, via a hitting criterion analogous to the Wiener test for Brownian motion.

Chapter 22

Functional Data Analysis and Infinite-Dimensional Statistics

22.1 Functional Data and Covariance Operators

Functional data analysis (FDA) studies data that are naturally functions $X_i : [0,1] \to \mathbb{R}$ (or more generally, random elements of a Hilbert space $\mathcal{H}$) rather than finite-dimensional vectors. The covariance operator replaces the covariance matrix.

Definition: Covariance Operator of a Random Element

Let X be a random element of a separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ with $\mathbb{E}[\|X\|^2] < \infty$. The **covariance operator** $C : \mathcal{H} \to \mathcal{H}$ is the bounded, self-adjoint, positive operator defined by

$$\langle f, Cg\rangle = \mathbb{E}[\langle f, X - \mathbb{E}X\rangle\langle g, X - \mathbb{E}X\rangle] = \operatorname{Cov}(\langle f, X\rangle, \langle g, X\rangle)$$

for all $f, g \in \mathcal{H}$. When $\mathcal{H} = L^2([0,1])$, C has the integral kernel $c(s,t) = \operatorname{Cov}(X(s), X(t))$.

Theorem: Karhunen–Loève Representation (FDA Form)

If C is a trace-class covariance operator with eigendecomposition $Ce_k = \lambda_k e_k$ ($\lambda_1 \geq \lambda_2 \geq \cdots > 0$, $\{e_k\}$ an ONB), then the centred process $\tilde{X} = X - \mathbb{E}[X]$ admits the L^2-convergent expansion

$$\tilde{X} = \sum_{k=1}^{\infty} \xi_k e_k, \qquad \xi_k = \langle \tilde{X}, e_k\rangle,$$

where $\mathbb{E}[\xi_k] = 0$, $\operatorname{Var}(\xi_k) = \lambda_k$, and $\operatorname{Cov}(\xi_j, \xi_k) = 0$ for $j \neq k$. If X is Gaussian, the ξ_k are mutually independent. The truncated expansion $\tilde{X}_K = \sum_{k=1}^{K} \xi_k e_k$ explains $\sum_{k=1}^{K} \lambda_k / \sum_{k} \lambda_k$ of the total variance.

22.2 Estimation of the Covariance Operator

Given n i.i.d. observations $X_1, \dots, X_n$ of a functional random variable, the natural estimator of C is the sample covariance operator.

Definition: Sample Covariance Operator

The **sample covariance operator** is

$$\hat{C}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{X}_i \otimes \tilde{X}_i), \qquad \tilde{X}_i = X_i - \bar{X}_n,$$

where $(f \otimes g)h = \langle g, h \rangle f$ is the rank-one operator formed by the tensor product of f and g, and $\bar{X}_n = n^{-1} \sum_i X_i$ is the sample mean. In integral kernel form, $\hat{c}_n(s,t) = (n-1)^{-1} \sum_i (X_i(s) - \bar{X}_n(s))(X_i(t) - \bar{X}_n(t))$.

Theorem: Convergence of Sample Covariance Operators (Dauxois–Pousse–Romain)

Let $X, X_1, X_2, \dots$ be i.i.d. in $L^2([0,1])$ with $\mathbb{E}[\|X\|^4] < \infty$ and covariance operator C. Then $\|\hat{C}_n - C\|_{\mathrm{HS}} \to 0$ almost surely, and $\sqrt{n}(\hat{C}_n - C) \xrightarrow{d} \mathcal{G}$ in Hilbert–Schmidt norm, where $\mathcal{G}$ is a Gaussian random operator with covariance determined by the fourth-order cumulants of X.

22.3 Functional Principal Components and FPCA

Functional principal components analysis (FPCA) extends PCA to the functional setting by extracting the directions in $\mathcal{H}$ of maximum variance.

Definition: Functional Principal Components

The **functional principal components** (FPCs) $v_1, v_2, \dots$ of a functional random variable X with covariance operator C are the orthonormal eigenfunctions of C:

$$C v_k = \lambda_k v_k, \quad \|v_k\| = 1, \quad \langle v_j, v_k \rangle = \mathbf{1}_{j=k},$$

ordered by decreasing eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. The k-th **FPC score** is $\xi_k = \langle X - \mathbb{E}X, v_k \rangle$.

Proposition: Optimal Reconstruction Property of FPCA

Among all rank-K approximations of the form $\hat{X}_K = \mathbb{E}[X] + \sum_{k=1}^{K} \langle X - \mathbb{E}X, \phi_k \rangle \phi_k$ with $\{\phi_k\}$ orthonormal, the FPCA approximation (using the FPCs $v_1, \dots, v_K$) minimises the mean squared reconstruction error:

$$\mathbb{E}\Big[\|X - \hat{X}_K\|^2\Big] = \sum_{k > K} \lambda_k,$$

which is minimised by taking $\phi_k = v_k$ (the FPCs).

Example: Longitudinal Growth Data

In the Berkeley Growth Study, $X_i(t)$ is the height (in cm) of child i at age $t \in [1, 18]$ years. FPCA of the $n = 93$ growth curves reveals that the first FPC v_1 (explaining $\approx 86\%$ of variance) corresponds to the overall height scale; the second FPC v_2 (explaining $\approx 11\%$) captures the timing of the adolescent growth spurt; and the third FPC v_3 (explaining $\approx 2\%$) reflects the intensity of the post-spurt deceleration. The FPC scores $\xi_{i1}, \xi_{i2}, \xi_{i3}$ then provide a three-dimensional summary of each child's growth trajectory, enabling clustering, regression on external covariates (sex, nutrition), and prediction of adult height from early measurements.

22.4 Functional Regression

Functional regression extends linear regression to the setting where the predictor and/or response is a function. We focus on the **function-on-scalar regression** and **scalar-on-function regression** models, both of which are natural GP models.

Definition: Scalar-on-Function Regression

In **scalar-on-function regression**, the response $Y_i \in \mathbb{R}$ is modelled as

$$Y_i = \alpha + \int_0^1 X_i(t)\,\beta(t)\,dt + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $\beta : [0, 1] \to \mathbb{R}$ is an unknown **coefficient function** to be estimated. In the FPCA representation, substituting $X_i(t) = \bar{X}(t) + \sum_k \xi_{ik} v_k(t)$ gives the equivalent model

$$Y_i = \mu + \sum_{k=1}^{\infty} \xi_{ik}\, b_k + \varepsilon_i, \quad b_k = \int_0^1 v_k(t)\beta(t)\,dt,$$

which is a linear model in the (infinitely many) FPC scores. Truncation to the first K components gives a finite-dimensional approximation estimated by ordinary least squares.

Theorem: Convergence Rate for Functional Linear Regression

Suppose $\beta \in \mathcal{H}_k$, the RKHS of the kernel $k(s,t) = \mathrm{Cov}(X(s), X(t)) = c(s,t)$, and the eigenvalues of C decay as $\lambda_k \asymp k^{-2\alpha}$ for some $\alpha > 1/2$. The truncated FPCA estimator $\hat{\beta}_K$ with $K \asymp n^{1/(2\alpha+1)}$ achieves:

$$\mathbb{E}\left[\|\hat{\beta}_K - \beta\|_{L^2}^2\right] = O\left(n^{-2\alpha/(2\alpha+1)}\right),$$

which is the minimax-optimal rate for estimating a function with α derivatives in L^2 from n noisy observations.

Chapter 23

Topological Data Analysis and Persistent Homology of Random Fields

23.1 Excursion Sets and Topological Complexity

The topology of the excursion set $\{x \in D : X(x) \geq u\}$ of a random field X at level u encodes global structural information about the field. The **Euler characteristic** (EC) of the excursion set is a fundamental topological invariant that admits a tractable expression in terms of the covariance structure.

Definition: Excursion Set and Euler Characteristic

For a smooth random field X on a domain $D \subset \mathbb{R}^d$ and a level $u \in \mathbb{R}$, the **excursion set** is $A_u(X, D) = \{x \in D : X(x) \geq u\}$. The **Euler characteristic** (EC) $\chi(A_u)$ is the alternating sum of Betti numbers: $\chi = \beta_0 - \beta_1 + \beta_2 - \cdots$, where β_k counts the number of k-dimensional holes (connected components, loops, voids, etc.) in A_u.

Theorem: Adler–Taylor Kinematic Formula for the EC

Let X be a centered, unit-variance, smooth stationary Gaussian field on $D \subset \mathbb{R}^d$ with covariance $k(h)$ and $k(0) = 1$. The expected Euler characteristic of the excursion set satisfies

$$\mathbb{E}[\chi(A_u(X, D))] = \sum_{j=0}^{d} \mathcal{L}_j(D)\, \rho_j(u),$$

where $\mathcal{L}_j(D)$ are the **Lipschitz–Killing curvatures** (intrinsic volumes) of D in the Riemannian metric induced by the covariance ($g_{ij} = -\partial_{s_i}\partial_{t_j} k(s-t)|_{s=t}$), and

$$\rho_j(u) = (2\pi)^{-(j+1)/2}\, H_j(u)\, e^{-u^2/2}$$

are the **EC densities**, with H_j the j-th Hermite polynomial. In particular, $\rho_0(u) = 1 - \Phi(u)$ (the Gaussian tail), $\rho_1(u) = (2\pi)^{-1} e^{-u^2/2}$, and $\rho_2(u) = (2\pi)^{-3/2} u e^{-u^2/2}$.

The Adler–Taylor formula is a central result in geometric probability: it connects the topological complexity of Gaussian excursion sets to the Riemannian geometry of the index set. It finds applications in cosmology (the topology of the cosmic microwave background), brain imaging (detecting activated brain regions), and materials science (the topology of porous media).

23.2 Persistent Homology and the Persistence Diagram

While the Euler characteristic captures a single number summarising topology at a fixed level, **persistent homology** tracks the birth and death of topological features as the level u varies, providing a multi-scale topological summary.

Definition: Persistence Diagram

For a random field X on D, consider the nested family of excursion sets $\{A_u\}_{u\in\mathbb{R}}$ (decreasing as u increases). Each connected component of A_u is born at some level b (the maximum of X over that component) and dies at level $d < b$ (when it merges with an older component or disappears). The **persistence diagram** $\mathrm{Dgm}(X)$ is the multiset of birth–death pairs $(b_i, d_i) \in \mathbb{R}^2$ with $b_i > d_i$, one for each topological feature (of any homological degree) across all levels.

Theorem: Stability of Persistence Diagrams

For two random fields X, Y on D, the **bottleneck distance** between their persistence diagrams satisfies

$$d_b(\mathrm{Dgm}(X), \mathrm{Dgm}(Y)) \le \|X - Y\|_{C^0(D)} = \sup_{x\in D} |X(x) - Y(x)|.$$

That is, the map $X \mapsto \mathrm{Dgm}(X)$ is Lipschitz with respect to the sup-norm and the bottleneck distance, implying that small perturbations of the field produce small perturbations of the persistence diagram.

Proof sketch. The stability theorem follows from the interleaving of filtered chain complexes. If $\|X - Y\|_\infty \le \varepsilon$, then $A_{u+\varepsilon}(X) \subseteq A_u(Y) \subseteq A_{u-\varepsilon}(X)$ for all u, which gives an ε-interleaving of the persistence modules $H_k(A_u(X))$ and $H_k(A_u(Y))$. By the algebraic stability theorem (Cohen-Steiner–Edelsbrunner–Harer, 2007), an ε-interleaving implies $d_b(\mathrm{Dgm}(X), \mathrm{Dgm}(Y)) \le \varepsilon$. □

23.3 Statistical Inference via the Persistence Diagram

The persistence diagram is a functional summary statistic of the random field, amenable to statistical hypothesis testing and estimation.

Definition: Wasserstein Distance for Persistence Diagrams

The p-**Wasserstein distance** between two persistence diagrams $\mathrm{Dgm}(X)$ and $\mathrm{Dgm}(Y)$ is

$$d_{W_p}(\mathrm{Dgm}(X), \mathrm{Dgm}(Y)) = \left(\inf_{\eta} \sum_{x \in \mathrm{Dgm}(X)} \|x - \eta(x)\|_\infty^p \right)^{1/p},$$

where the infimum is over all bijections $\eta : \mathrm{Dgm}(X) \to \mathrm{Dgm}(Y)$, allowing points to be matched to the diagonal $\{(b, b) : b \in \mathbb{R}\}$ (representing the birth and immediate death of trivial features) at cost $\mathrm{pers}(x)/2$ where $\mathrm{pers}(x) = b_x - d_x$ is the persistence.

Theorem: Central Limit Theorem for Persistent Betti Numbers (Hiraoka et al.)

For a stationary Gaussian field X on $[0,1]^d$ and a fixed level u, the random variable $\beta_k(A_u)$ (the k-th Betti number of the excursion set) satisfies, as the field is observed on a growing domain $[0, n]^d$ with $n \to \infty$:

$$\frac{\beta_k(A_u([0,n]^d)) - n^d \mu_k(u)}{\sqrt{n^d}} \xrightarrow{d} \mathcal{N}(0, \sigma_k^2(u)),$$

where $\mu_k(u) = \mathbb{E}[\beta_k(A_u([0,1]^d))]$ is the expected Betti number density and $\sigma_k^2(u)$ is the asymptotic variance, expressible via the covariance structure of the field.

Example: Detecting Signals in Noise via Topology

In cosmology, the temperature fluctuations of the CMB are modelled as a nearly Gaussian random field on the sphere S^2. The persistence diagram of the excursion sets of the CMB temperature field has been compared to those of simulated Gaussian fields, and statistically significant anomalies in the Wasserstein distance provide evidence of non-Gaussianity beyond what is detectable by the power spectrum alone. The topological data analysis of the CMB has identified the "cold spot" anomaly as a particularly persistent topological feature (large persistence = large $|b - d|$), corresponding to an unusually large and deep negative temperature fluctuation.

Chapter 24

Gaussian Processes in Reinforcement Learning and Bayesian Optimisation

24.1 Bayesian Optimisation

Bayesian optimisation (BO) is a framework for the global optimisation of expensive, noisy, black-box functions $f : \mathcal{X} \to \mathbb{R}$. It places a GP prior on f and uses the posterior to guide the selection of evaluation points, balancing exploration (reducing uncertainty in unexplored regions) and exploitation (evaluating near the current best point).

Definition: Bayesian Optimisation with GP Prior

Let $f : \mathcal{X} \to \mathbb{R}$ be the unknown function to optimise, with GP prior $f \sim \mathcal{GP}(0, k)$. Given evaluations $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ with $y_i = f(x_i) + \varepsilon_i$, the **Bayesian optimisation** algorithm selects the next evaluation point x_{n+1} by maximising an **acquisition function** $\alpha_n : \mathcal{X} \to \mathbb{R}$ that summarises the desirability of each candidate point under the posterior $f \mid \mathcal{D}_n \sim \mathcal{GP}(\mu_n, k_n)$.

Definition: Expected Improvement Acquisition Function

The **Expected Improvement** (EI) acquisition function is

$$\mathrm{EI}(x; \mathcal{D}_n) = \mathbb{E}[\max(f(x) - f_n^*, 0) \mid \mathcal{D}_n],$$

where $f_n^* = \max_{i=1}^n y_i$ is the current best observation. Under the GP posterior $f(x) \mid \mathcal{D}_n \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$, this evaluates to

$$\mathrm{EI}(x; \mathcal{D}_n) = (\mu_n(x) - f_n^*)\Phi(Z) + \sigma_n(x)\phi(Z),$$

where $Z = (\mu_n(x) - f_n^*)/\sigma_n(x)$, and ϕ, Φ are the standard Gaussian density and CDF.

Theorem: Cumulative Regret Bound for GP-UCB

The **GP-UCB** acquisition function $\alpha_n(x) = \mu_n(x) + \sqrt{\beta_n}\,\sigma_n(x)$ with $\beta_n = 2\log(|\mathcal{X}|n^2\pi^2/6\delta)$ achieves the cumulative regret bound:

$$R_n = \sum_{t=1}^{n}(f(x^*) - f(x_t)) \leq \sqrt{C_1 n \beta_n \gamma_n}$$

with probability at least $1-\delta$, where $x^* = \arg\max f$, $C_1 = 8/\log(1+\sigma_n^{-2})$, and $\gamma_n = \max_{A\subseteq\mathcal{X},|A|=n} I(f_A; y_A)$ is the **maximum information gain** (the mutual information between n function values and corresponding noisy observations under the GP prior).

Proposition: Maximum Information Gain for Matérn Kernels

For the Matérn-ν kernel on $[0,1]^d$ and n evaluation points:

$$\gamma_n = O\big((\log n)^{d+1}\big) \quad (\nu = \infty,\ \text{SE kernel}), \qquad \gamma_n = O\big(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\big) \quad (\nu < \infty).$$

Substituting into the regret bound gives: for the SE kernel, $R_n = O(\sqrt{n}(\log n)^{(d+1)/2+1})$, achieving sublinear regret; for Matérn-ν, $R_n = O(n^{(d+\nu(d+1))/(2\nu+d)})$, which is sublinear when $\nu > d$.

24.2 Thompson Sampling and the Exploration–Exploitation Trade-off

An alternative to acquisition-function-based BO is **Thompson sampling** (TS), which selects each evaluation point by sampling a function from the GP posterior and returning its maximiser.

Definition: GP Thompson Sampling

At round $n+1$, given $\mathcal{D}_n$:

1. Sample $\tilde{f} \sim \mathcal{GP}(\mu_n, k_n)$ (a draw from the posterior);
2. Set $x_{n+1} = \arg\max_{x\in\mathcal{X}} \tilde{f}(x)$.

The next observation is $y_{n+1} = f(x_{n+1}) + \varepsilon_{n+1}$, and the posterior is updated to $\mathcal{GP}(\mu_{n+1}, k_{n+1})$.

Thompson sampling achieves the same regret bounds as GP-UCB (up to logarithmic factors) while being easier to implement in distributed settings: multiple agents can each independently sample $\tilde{f}$ from the posterior and evaluate at their own maximiser, without the need to coordinate their acquisition function evaluations.

24.3 Multi-Fidelity and Multi-Task Gaussian Processes

In many scientific applications, evaluations of the target function f are available at multiple fidelities (accuracy levels), with lower-fidelity evaluations being cheaper but noisier. Multi-fidelity GPs model the relationship between fidelity levels to leverage cheap low-fidelity data to guide the selection of expensive high-fidelity evaluations.

Definition: Auto-Regressive Multi-Fidelity GP

The **auto-regressive (AR1) multi-fidelity model** (Kennedy–O'Hagan, 2000) posits M fidelity levels with

$$f_m(x) = \rho_{m-1}\, f_{m-1}(x) + \delta_m(x), \quad m = 1, \dots, M,$$

where $\delta_m \sim \mathcal{GP}(0, k_m)$ are independent GP discrepancies, f_0 is a GP prior for the lowest fidelity, and ρ_m are scalar scaling coefficients. The highest-fidelity output is f_M. Given data $\{(x_i, y_i^{(m)})\}$ at various fidelities, the posterior of f_M at new points x^* is computed by propagating the conditional means and covariances through the AR1 recursion.

Remark: Multi-Task GP and Transfer Learning

When the outputs $\{f_1, \dots, f_M\}$ are not a fidelity hierarchy but rather M related tasks (e.g., the aerodynamic drag on M different aircraft designs), the **multi-task GP** (Bonilla–Chai–Williams, 2007) models the joint distribution as

$$\begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix} \sim \mathcal{GP}\big(\mathbf{0},\, LL^{\top} \otimes k(x, x')\big)\,,$$

where L is a lower-triangular matrix (Cholesky factor of the inter-task covariance) and k is a shared kernel on the input space. This allows transfer of information between tasks: data on task m informs the posterior of task m' through the inter-task covariance $LL^{\top}_{mm'}$.

Chapter 25

Unified Perspectives: Kernels, Information, and Spatial Geometry

25.1 The Kernel as the Unifying Thread

Throughout this volume, the kernel—a positive-definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$—has appeared in three distinct but deeply connected roles: as the covariance function of a Gaussian random field, as the reproducing kernel of a Hilbert space of functions, and as the Green's function of an elliptic operator on a domain or graph. This final chapter draws these threads together, making explicit the connections that run beneath the surface of the preceding chapters.

Theorem: Equivalence of GP, RKHS, and Operator Perspectives

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric, positive-definite kernel on a metric space $(\mathcal{X}, d)$. The following three objects are in natural bijection:

(i) A centered Gaussian process $\{f(x) : x \in \mathcal{X}\}$ with covariance $\mathbb{E}[f(x)f(y)] = k(x, y)$;

(ii) The reproducing kernel Hilbert space $\mathcal{H}_k$ with inner product $\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}_k} = k(x, y)$;

(iii) The range of the integral operator $T_k : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$, $(T_k f)(x) = \int k(x, y) f(y)\, d\mu(y)$, equipped with the norm $\|T_k^{1/2} f\|_{L^2}$.

The GP posterior given observations $(x_1, y_1), \dots, (x_n, y_n)$ has mean $\hat{f} = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ (the minimum-norm interpolant in $\mathcal{H}_k$), and this is also the MAP estimator under the GP prior with an $\mathcal{H}_k$ regularisation penalty.

This unification has immediate practical consequences. The three perspectives correspond to three computational strategies: sampling from the GP (generative modelling), solving a quadratic program in $\mathcal{H}_k$ (kriging, support vector machines), and solving a linear system involving T_k (Nyström approximation, random Fourier features). Each strategy has different computational complexity and approximation guarantees, and the choice depends on the application.

Remark: The RKHS as a Bayesian Prior

The RKHS norm $\|f\|_{\mathcal{H}_k}^2$ plays the role of a (log-)prior density: functions with small RKHS norm are "smooth" relative to the kernel and are a priori more probable. The Bayesian GP model with $f \sim \mathcal{GP}(0, k)$ assigns prior density $\propto \exp(-\frac{1}{2}\|f\|_{\mathcal{H}_k}^2)$ (in the formal sense of a Gaussian measure on the function space). The regularised least-squares estimator $\hat{f}_\lambda = \arg\min_{f \in \mathcal{H}_k} \sum_i (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_k}^2$ is thus the MAP estimator under the GP prior with observation noise $\sigma^2 = \lambda$, establishing the complete duality between kernel methods and Bayesian nonparametric regression.

25.2 Information-Theoretic Bounds and the Fisher Information Geometry of Random Fields

The statistical analysis of random fields—estimation of covariance parameters, hypothesis testing for stationarity, model selection—relies on the Fisher information matrix and Cramér–Rao lower bounds. For Gaussian random fields, these have explicit expressions in terms of the kernel.

Proposition: Fisher Information for Gaussian Random Fields

Let $\{X_\theta(s) : s \in D\}$ be a Gaussian random field with zero mean and covariance kernel $k_\theta(s,t)$, observed at n points $s_1, \dots, s_n$. The Fisher information matrix for $\theta \in \mathbb{R}^p$ is

$$\mathcal{I}_{jl}(\theta) = \frac{1}{2} \operatorname{tr}\left(K_\theta^{-1} \frac{\partial K_\theta}{\partial \theta_j} K_\theta^{-1} \frac{\partial K_\theta}{\partial \theta_l} \right),$$

where $K_\theta = (k_\theta(s_i, s_j))_{i,j=1}^n$ is the $n \times n$ kernel matrix. The Cramér–Rao bound states that any unbiased estimator $\hat{\theta}$ satisfies $\operatorname{Cov}(\hat{\theta}) \succeq \mathcal{I}(\theta)^{-1}$.

Example: Fisher Information for the Matérn-ν Kernel

For the isotropic Matérn-ν kernel on $\mathbb{R}^d$ with range ℓ and variance σ^2:

$$k_{\sigma^2,\ell}(r) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, r}{\ell} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}\, r}{\ell} \right),$$

the partial derivatives $\partial K / \partial \sigma^2 = K/\sigma^2$ and $\partial K / \partial \ell$ have explicit forms via the derivative of the modified Bessel function K_ν. In the increasing-domain asymptotic regime (the observation domain grows as $n \to \infty$ at fixed spacing), the Fisher information for (σ^2, ℓ) grows as $O(n)$, yielding $\sqrt{n}$-consistent estimators of both parameters. In the fixed-domain asymptotic regime (the domain is fixed and the spacing shrinks to 0), only the *microergodic parameter* $\sigma^2 / \ell^{2\nu}$ is consistently estimable from the data; σ^2 and ℓ individually are not identifiable in the limit.

The phenomenon of microergodicity—whereby only certain combinations of kernel parameters are estimable from infill observations—is a fundamental constraint on spatial statistics that has no finite-dimensional analogue. It arises because, in the fixed-domain setting, all observations lie within a compact region, and the local behaviour of the kernel (captured by $\sigma^2/\ell^{2\nu}$ for the Matérn family) is determined by the data, while the large-scale structure (captured by ℓ and σ^2 individually) is not.

25.3 Spatial Geometry and the Intrinsic Random Field

Stationary random fields require a homogeneous background (typically $\mathbb{R}^d$ or a compact Lie group). When the physical domain has non-trivial geometry—a curved surface, a network, a hierarchical tree—stationarity must be replaced by a weaker but still tractable concept.

Definition: Intrinsic Random Field of Order k (IRF-k)

A random field $\{Z(x) : x \in \mathbb{R}^d\}$ is an **intrinsic random field of order** k (IRF-k) if, for every vector of weights $\{\lambda_i\}_{i=1}^n$ satisfying $\sum_i \lambda_i x_i^j = 0$ for $j = 0, 1, \dots, k$ (a *generalised increment* of order $k+1$), the linear combination $\sum_i \lambda_i Z(x_i)$ is a second-order stationary random variable whose variance depends only on the differences $\{x_i - x_j\}$. The *generalised covariance function* $K_k(h)$ satisfies $\mathrm{Var}(\sum_i \lambda_i Z(x_i)) = -\sum_{i,j} \lambda_i \lambda_j K_k(x_i - x_j)$ for all such $\{\lambda_i\}$.

Theorem: Characterisation of Generalised Covariances (Matheron)

The class of generalised covariance functions of order k on $\mathbb{R}^d$ consists of all conditionally negative-definite functions of the form

$$K_k(h) = (-1)^{k+1} \int_{\mathbb{R}^d} \frac{1 - \cos(\omega \cdot h)}{\|\omega\|^{2k+d}} \, \hat{\mu}(d\omega) + P_{2k}(h),$$

where $\hat{\mu}$ is a non-negative spectral measure and P_{2k} is a polynomial of degree at most $2k$. In the isotropic case on $\mathbb{R}^1$, the polynomial generalised covariances of order k are $K_k(h) = a_0 + a_1|h| + \cdots + a_k|h|^{2k-1}$ with alternating signs for the a_j.

Universal kriging with IRF-k generalised covariances is the natural geostatistical framework for data exhibiting spatial trends of polynomial degree k: the trend is not modelled parametrically but is absorbed into the class of allowable interpolation weights, and the kriging predictor is the linear unbiased estimator with minimum variance within this class.

25.4 From Random Fields to Random Measures: The Measure-Valued Perspective

The measure-valued perspective unifies random fields, point processes, and the Gaussian free field within a single framework: all three are random elements of a space of measures (or distributions) on the domain $\mathcal{X}$.

Definition: Random Measure

A **random measure** on a locally compact, separable metrizable space $(\mathcal{X}, \mathcal{B})$ is a measurable map $\mu : \Omega \to \mathcal{M}(\mathcal{X})$, where $\mathcal{M}(\mathcal{X})$ is the space of locally finite non-negative Borel measures on $\mathcal{X}$, equipped with the vague topology (the weak-* topology against compactly supported continuous functions). The law of μ is a probability measure on $\mathcal{M}(\mathcal{X})$.

Random measures subsume the principal objects of this volume:

(i) A **point process** $\Phi = \sum_i \delta_{x_i}$ is an integer-valued random measure;

(ii) A **Cox process** directed by a random intensity Λ is a doubly stochastic random measure: conditionally on Λ, it is a Poisson random measure with intensity Λ;

(iii) The **Gaussian free field** Φ on D is a random distribution (signed measure in the distributional sense), taking values in the dual Sobolev space $H^{-1}(D)$ for $d = 2$;

(iv) A **spatial Gibbs measure** is a probability measure on the space of locally finite configurations, characterising a random measure through its DLR equations.

Theorem: Characterisation of Infinitely Divisible Random Measures (Kingman)

A random measure μ on $\mathcal{X}$ is **infinitely divisible** (i.e., for every n, $\mu \overset{d}{=} \mu_1 + \cdots + \mu_n$ for i.i.d. μ_i) if and only if its Laplace functional has the form

$$\log \mathcal{L}_\mu(f) = -\int_{\mathcal{X}} f(x)\,\alpha(dx) - \int_{\mathcal{M}_0(\mathcal{X})} \left(1 - e^{-\nu(f)}\right) Q(d\nu),$$

where α is a non-negative measure (the deterministic part), $\nu(f) = \int f\, d\nu$, and Q is a sigma-finite measure on $\mathcal{M}_0(\mathcal{X}) = \mathcal{M}(\mathcal{X}) \setminus \{0\}$ (the Lévy–Khintchine measure or *KLM measure*) satisfying $\int \min(1, \nu(\mathcal{X}))\, Q(d\nu) < \infty$. The Poisson process corresponds to $Q = \int_{\mathcal{X}} \delta_{\delta_x}\, \lambda(dx)$ (unit-mass atoms), and Cox processes arise when Q is concentrated on non-atomic measures.

25.5 Approximation Theory for Gaussian Processes: Nyström, Inducing Points, and Scalable Inference

For large datasets, exact GP inference requiring $O(n^3)$ computation via Cholesky factorisation of the kernel matrix becomes infeasible. A systematic theory of approxima-

tions connects numerical linear algebra, approximation theory in RKHSs, and information theory.

Definition: Nyström Approximation and Inducing Points

Given n observations $\{x_i\}$ and $m \ll n$ **inducing points** $\{z_j\}_{j=1}^m$, the **Nyström approximation** to the kernel matrix K_{nn} is

$$\tilde{K}_{nn} = K_{nm} K_{mm}^{-1} K_{mn},$$

where $K_{nm} = (k(x_i, z_j))$ is the $n \times m$ cross-kernel matrix and $K_{mm} = (k(z_i, z_j))$ is the $m \times m$ inducing-point kernel matrix. The GP posterior under the Nyström approximation reduces to $O(nm^2)$ computation (the dominant cost is the $n \times m$ matrix multiplication) and $O(nm)$ storage.

Theorem: Approximation Error of the Nyström Method

Let $T_k^{(n)}$ be the empirical kernel operator $(T_k^{(n)} f)(x) = n^{-1} \sum_i k(x, x_i) f(x_i)$ and let $T_k^{(m)}$ be its Nyström approximation based on m inducing points drawn uniformly from $\{x_i\}$. Then with probability at least $1 - \delta$:

$$\left\| T_k^{(n)} - T_k^{(m)} \right\|_F \le \left\| T_k^{(n)} \right\|_F \sqrt{\frac{4k \log(2/\delta)}{m}},$$

where k is the rank of the top-k eigenspace of $T_k^{(n)}$. In particular, choosing $m = O(k \log(1/\delta)/\varepsilon^2)$ inducing points achieves ε-relative error in Frobenius norm.

Example: Optimal Inducing Point Selection via Information Gain

The **variational inducing-point method** (Titsias, 2009) selects the inducing points $\{z_j\}$ by maximising the evidence lower bound (ELBO)

$$\mathcal{L}(\{z_j\}, \mathbf{m}, S) = \sum_{i=1}^{n} \mathbb{E}_{q(f_i)}[\log p(y_i \mid f_i)] - \mathrm{KL}(q(\mathbf{u}) \| p(\mathbf{u})),$$

where $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, S)$ is the variational posterior over the inducing outputs $\mathbf{u} = f(\{z_j\})$, and $q(f_i) = \int p(f_i \mid \mathbf{u}) q(\mathbf{u})\, d\mathbf{u}$ is the implied marginal. For Gaussian likelihoods, the ELBO is available in closed form and equals the log marginal likelihood of a projected process model, minus a trace term $\frac{1}{2\sigma^2} \operatorname{tr}(K_{nn} - \tilde{K}_{nn})$ that penalises the approximation error. The optimal inducing points maximise this ELBO, and for a fixed m, the greedy selection of z_j one at a time is equivalent to maximising the information gain $I(f; \mathbf{u})$ at each step.

25.6 Connections to Deep Learning: Neural Tangent Kernels and Feature Learning

The final frontier connecting random field theory to contemporary machine learning is the analysis of deep neural networks in the infinite-width limit, where they become Gaussian processes with specific kernels.

Theorem: Neural Tangent Kernel (Jacot–Gabriel–Hongler, 2018)

Consider a fully connected network $f_\theta : \mathbb{R}^{d_0} \to \mathbb{R}$ of depth L with weights θ, trained by gradient descent on the loss $\mathcal{L}(\theta) = \sum_i \ell(f_\theta(x_i), y_i)$. As the widths $n_1, \dots, n_{L-1} \to \infty$ simultaneously, the **neural tangent kernel**

$$K^{\mathrm{NTK}}(x, x') = \mathbb{E}_\theta[\nabla_\theta f_\theta(x) \cdot \nabla_\theta f_\theta(x')]$$

converges to a deterministic limiting kernel $K^\infty(x, x')$, which satisfies the recursion

$$\begin{aligned} K^{(0)}(x, x') &= x \cdot x', \\ \Sigma^{(h)}(x, x') &= \mathbb{E}_{(u,v)\sim\mathcal{N}(0, K^{(h-1)}(x,x'))}[\sigma(u)\sigma(v)], \\ K^{(h)}(x, x') &= \Sigma^{(h)}(x, x') + K^{(h-1)}(x, x') \cdot \dot{\Sigma}^{(h)}(x, x'), \end{aligned}$$

where σ is the activation function and $\dot{\Sigma}^{(h)}$ involves the derivative σ'. Under gradient flow, the network outputs evolve as $\dot{f}_\theta(x) = -K^\infty \cdot \nabla_f \mathcal{L}$, making the dynamics exactly linear in the function space and the solution computable as a kernel regression with kernel K^∞.

Remark: Feature Learning Beyond the NTK Regime

The NTK regime, in which weights move only infinitesimally from their initialisation, is now understood to miss an important aspect of deep learning: **feature learning**, the ability of the network to adapt its internal representations to the data. In the NTK regime, the kernel K^∞ is fixed and the network is equivalent to a kernel method; in the feature-learning regime (finite width, large learning rate, or mean-field parameterisation), the kernel itself evolves during training. The tensor program framework (Yang, 2019–2021) and the μP parameterisation provide a systematic theory of feature learning, extending the infinite-width analysis to regimes where the kernel can adapt. This remains an active research frontier, with connections to information-theoretic analyses of the Fisher information geometry of the loss landscape.

Proposition: GP Posterior and NTK Regression are Equivalent

For a fully connected network in the infinite-width NTK limit, trained on data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ by gradient flow to convergence with squared loss, the network output $f_{\theta_\infty}(x^*)$ at a test point x^* equals the GP posterior mean under the GP prior

$f \sim \mathcal{GP}(0, K^\infty)$ with observation noise $\sigma^2 \to 0$:

$$f_{\theta_\infty}(x^*) = \mathbf{k}_*^\top K^{-1} \mathbf{y},$$

where $\mathbf{k}_* = (K^\infty(x^*, x_1), \ldots, K^\infty(x^*, x_n))^\top$ and $K = (K^\infty(x_i, x_j))_{ij}$. The posterior variance $K^\infty(x^*, x^*) - \mathbf{k}_*^\top K^{-1} \mathbf{k}_*$ quantifies the epistemic uncertainty not captured by the point estimate.

Appendix A: Notation and Conventions

Throughout this volume, we use the following notational conventions.

Sets and spaces. $\mathbb{R}^d$ denotes the d-dimensional Euclidean space with inner product $\langle x, y\rangle = \sum_{i=1}^{d} x_i y_i$ and norm $\|x\| = \sqrt{\langle x, x\rangle}$. $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra of a topological space $\mathcal{X}$. $\mathbf{N}(\mathcal{X})$ is the space of locally finite counting measures on $\mathcal{X}$.

Probability. $\mathbb{P}$ always denotes a probability measure; $\mathbb{E}$ expectation; Var variance; Cov covariance. δ_x is the Dirac measure at x. $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian distribution with mean μ and variance σ^2; $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is the multivariate Gaussian. $\mathcal{GP}(m, k)$ is a Gaussian process with mean function m and covariance kernel k.

Kernels and operators. K denotes both a kernel function and the Gram matrix $(k(s_i, s_j))$ in context. $\mathcal{H}_k$ is the RKHS of k. T_k is the integral operator associated to k. $G_D(x, y)$ is the Green's function of $-\Delta$ on domain D.

Point processes. $\Phi \sim \mathrm{PPP}(\Lambda)$ is a Poisson point process with intensity measure Λ. $\mathcal{L}_\Phi$ is the Laplace functional. $K(r)$ is Ripley's K-function. $g(r)$ is the pair correlation function.

Appendix B: Summary of Main Theorems

The following table summarises the principal theorems of this volume and the chapters in which they are proved.

Theorem	Chapter	Key Implication
Moore–Aronszajn	1	RKHS uniquely determined by kernel
Bochner	1	Spectral representation of stationary covariances
Kolmogorov Extension	1	Existence of random fields from consistent fidi's
Mercer	1	Kernel eigenfunction expansion
Schoenberg	8	Characterisation of universal kernels
Kolmogorov–Chentsov	3	Hölder continuity of sample paths
Fernique	4	Gaussian tails for field suprema
Dudley Entropy	4	Metric entropy bound for mean supremum
Borell–TIS	4	Concentration inequality for field maxima
Existence of PPP	6	PPP exists for all locally finite intensities
Campbell (CLM)	6	First/second moment formula for PPP
Slivnyak	6	Palm = law for PPP (characterisation)
Peierls	9	Phase transition in $\mathbb{Z}^2$ Ising model
Lee–Yang	9	Zeros of partition function on unit circle

FKG Inequality	9	Positive correlations for ferromagnets
Macchi–Soshnikov	10	Characterisation of DPPs via eigenvalues
Optimality of Kriging	11	BLUP characterisation
GP Posterior Contraction	12	Minimax-optimal rate for GP regression
NNGP Correspondence	14	Infinite-width NNs are GPs
Matérn SPDE	15	Matérn fields as SPDE solutions

Appendix C: Bibliographic Notes

Chapters 1–3. The theory of reproducing kernel Hilbert spaces originates with Aronszajn (1950) and Mercer (1909). Kolmogorov's extension theorem is in Kolmogorov (1933). Bochner's theorem in its general form is due to Bochner (1933). Schoenberg's characterisation of radial positive-definite functions appears in Schoenberg (1938). The Kolmogorov–Chentsov theorem for random fields is proved in detail in Adler–Taylor (2007), *Random Fields and Geometry*.

Chapters 4–5. Fernique's theorem is from Fernique (1975). The Dudley metric entropy bound is in Dudley (1967). The Borell–TIS inequality appears in Borell (1975) and Tsirelson–Ibragimov–Sudakov (1976). The Gaussian free field is treated in depth in Sheffield (2007) and Werner–Powell (2021). The Duplantier–Sheffield theorem on LQG is from Duplantier–Sheffield (2011).

Chapters 6–8. The foundational treatment of point processes is Daley–Vere-Jones (2003, 2008), *An Introduction to the Theory of Point Processes*, Vols. I–II. Campbell's theorem: Campbell (1909). Slivnyak's theorem: Slivnyak (1962). The CLM formula: Mecke (1967). Log-Gaussian Cox processes: Møller–Syversveen–Waagepetersen (1998). Palm calculus: Baccelli–Blaszczyszyn (2009).

Chapters 9. The Peierls argument is in Peierls (1936). The exact solution of the 2D Ising model is Onsager (1944). The Lee–Yang theorem is in Lee–Yang (1952). The FKG inequality is from Fortuin–Kasteleyn–Ginibre (1971). A comprehensive treatment is Friedli–Velenik (2017), *Statistical Mechanics of Lattice Systems*.

Chapters 10. Determinantal point processes are introduced in Macchi (1975) and developed in Soshnikov (2000). The GUE eigenvalue process and sine kernel are in Mehta (2004), *Random Matrices*. Sampling algorithms for DPPs: Kulesza–Taskar (2012). Zeros of Gaussian analytic functions: Hough–Krishnapur–Peres–Virág (2009).

Chapters 11–13. Kriging is developed in Matheron (1963, 1971). The universal kriging theorem (BLUP) is in Goldberger (1962). REML: Patterson–Thompson (1971). GP regression from the machine learning perspective: Rasmussen–Williams (2006), *Gaussian Processes for Machine Learning*. Posterior contraction rates: Van der Vaart–Van Zanten (2008, 2011). The SKI approximation is in Wilson–Nickisch (2015).

Chapters 14–15. The NNGP correspondence is established in Neal (1996) for single hidden layer networks and in Matthews et al. (2018), Garriga-Alonso et al. (2019) for deep networks. The SPDE approach to Matérn fields: Whittle (1954), Lindgren–Rue–Lindström (2011). Stochastic heat equations: Walsh (1986), Da Prato–Zabczyk (1992).